5-2018

Homeolog Gene Regulation in the Developing Allotetraploid Frog, Xenopus laevis

Ronald Cutler

Homeolog Gene Regulation in the Developing Allotetraploid Frog, *Xenopus laevis*

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Biology from the College of William and Mary

By

Ronald R. Cutler

Accepted for _____

_____

Margaret S. Saha, Ph.D., Director

_____

Joshua R. Puzey, Ph.D.

_____

Alban Guillaumet, Ph.D.

_____

Peter Kemper, Ph.D.

Acknowledgements

I would first a foremost like to give a huge thanks to my advisor Margaret Saha. Through her mentorship, I have cultivated a passion for the process of scientific inquiry to learn more about the world and its many mysterious processes. I have learned from her to remain skeptical and question any statement or result that I encounter.

My analysis could not have been possible if it was not for the wonderful members, past and present, of the Saha lab, who have put in countless hours of work to collect much of the data used in this work data as well as guide me in understanding the many techniques of molecular biology despite my inexperience in the wet-lab. This includes Lyuba, Andy, Chen, Mark, Sudip and Charith for their contributions in the Notch and Anterior-Posterior Rotation projects. I have also had incredible support from the bioinformatics group in the lab which includes my original mentor Caroline, LeAnn, and Josh who were an essential part of conducting the analysis in this work.

I would like to thank the members of my committee, Dr. Guillamet, Dr. Kemper, and Dr. Puzey for their valuable input and support of my thesis. I would also like to thank the financial support from the Howard Hughes Medical Institute as well at the National Institutes of Health which fund my research.

Finally, the support from my friends and family have been an integral part of keeping me grounded during my time in college and especially during the time working on my thesis. I especially want to thank my girlfriend Morgan, who has always been a shoulder to lean on.

Table of Contents

Abstract

The high incidence of polyploidy in the *Xenopus* Genus suggest that there might be advantages to this phenomenon. A recent allotetraploidy in the lineage of the model organism, *Xenopus laevis*, has created homeologous copies of genes that are estimated to make up ~46% of the genome. In this study, we took a global approach to study the gene expression patterns of homeologous gene copies during normal embryonic development at the mid-gastrula, mid-neurula, and late tailbud stages. Using RNA-sequencing data, we have characterized the high variance of homeolog gene expression which can be used to group homeologs into functionally distinct categories. Moreover, we have found that while there is an expression bias towards L homeologs, expression variance is biased towards S homeologs. We further characterized homeolog expression patterns during perturbations to embryonic development at the mid-neurula, early tailbud, and late tailbud stages where we found significant changes in homeolog expression bias that may be involved in the response and compensation from these perturbations. Our results suggest that: 1) genetic redundancy might confer advantages by allowing homeologous genes to diversify in gene function which is associated with highly variable homeolog expression patterns; 2) homeologs are highly involved in the response to perturbations during development and may provide a genetic buffer which may play an important role in the compensatory process.

Background

Polyploidy

A polyploid organism is one that has more than two sets of homologous chromosomes. As humans, we are familiar with our diploid (2N) genome which contains exactly two sets of homologous chromosomes, excluding the sex chromosomes. However, throughout nature polyploidy is quite common; known to occur commonly in plants, insects, fish, reptiles and amphibians (Otto, 2007). For example, wheat, a worldwide staple food, has 6 chromosome pairs (hexaploid) (Yang et al., 2015) while the Atlantic salmon is known to have 4 chromosome pairs (tetraploidy) (Lien et al., 2016).

The first discoveries of polyploidy stem from plant cytogenetics in the early twentieth century. While studying *Solanum* nigrum, Winkler (Winkler, 1916) discovered that cutting out explants from the stem had regenerated cells that were tetraploid. This discovery led him to first propose the term polyploidy in 1916 (Grant, 1971).

Since then, the implications of polyploidy have revealed mechanisms of evolution of novel gene function, adaptation, and speciation (Otto, 2007; Session et al., 2016; Wendel, 2015). Genomes that are duplicated due to polyploidy experience relaxed constraints on gene function. Pleiotropic genes which perform many functions now have copies which provide extra "degrees of freedom." This can allow for new specialized functions to evolve in these copies through mutation or change in spatial and/or temporal gene expression. On the other hand, subfunctionalization can result in the loss of this gene copy from the respective genome due to silencing and detrimental mutations. Thus, polyploidy provides genomic variation that can drives the evolution of a species.

Allopolyploidy

Two mechanisms of becoming polyploidy exist, autopolyploidy and allopolyploidy. Autopolyploidy refers to a whole genome duplication within one species that results in a duplicated set of identical chromosomes. In the case of a 2N (N = number of chromosome sets) organism, this *intra*specific genome doubling creates a 4N organism in which the duplicated genes and chromosomes are referred to as 'Ohnologs'. In contrast, allopolyploidy refers to a whole genome duplication due to a *inter*specific hybridization event between two closely related 2N species which create a 4N organism in which the duplicated genes and chromosomes are referred to as Homeologs (Glover, Redestig, & Dessimoz, 2016). The important difference between the two mechanisms is whether the extra set of chromosomes originates from the same species (auto) or two progenitor species (allo).

Distinction between auto and allopolyploids is key when studying polyploid organisms. Inference of an allopolyploid organism comes down to first distinguishing between the different homeologous subgenomes on the chromosomal level and then identifying similar homeologous genes found on opposing subgenomes on the gene level. However, to remain distinct from an autopolyploid, it must be shown that the subgenomes in an allopolyploid originated between two separate species.

In some cases, genomes of extant progenitor species – closely related species that hybridized to form an allopolyploid - can be used directly to achieve the auto and allopolyploidy distinction. First, sequences from a candidate allopolyploid can be separated into their respective subgenomes. Then orthology, sequence similarity and synteny, between each of the progenitor species (constrained to their corresponding subgenome) can be used to infer allopolyploidy (Cox et al., 2014).

In cases where the progenitor species are no longer extant, other computational methods must be used. Since each subgenome within an allopolyploid originated from a separate species, bivalent chromosome associations within each subgenome during metaphase 1 of meiosis can be used to infer allopolyploidy (Glover et al., 2016; Mason & Pires, 2015). However, these pairing behaviors are not exact and inconsistent among polyploids. Taking advantage of transposon sequences that are hypothesized to have independently originated in each of the progenitor species, Session *et al*. used distinct mariner and harbinger transposon sequences that were specific to each subgenome respectively to confirm allotetraploidy in *Xenopus laevis*.

Allopolyploidy Xenopus laevis

An allopolyploidy event between two progenitor African clawed frog species *Xenopus (L)* and *Xenopus (S)* is hypothesized to have occurred between 17-18 million years ago (MYA) (Session et al., 2016). This inference was made by comparison of the activity of transposable elements specific to each subgenome, namely mariner and harbinger transposon sequences as mentioned above. Comparing this subgenome specific activity to transposable element activity that is now uniform across both subgenomes in the current allotetraploid, an accurate estimate of the allotetraploid event between the progenitors species was made. This same method was used in another related allotetraploid frog *Xenopus borealis,* which returned a similar estimate that agrees with the original estimate of the allotetraploid event and offers further validation (Session et al., 2016).

The mariner and harbinger sequences also contributed to the definitive identification of the two diploid subgenomes of *X. laevis* denoted L and S for long and short respectively. These names refer to the to the relative lengths of the homeologous chromosomes where the S

chromosomes assembled sequence are on average 17.3% shorter than the L chromosomes. This discrepancy is due to asymmetric gene loss between the homeologous genomes that has affected the S subgenome (31.5%) significantly more than the L subgenome (8.3%). Despite the differences in gene loss between the homeologous chromosomes, *X. laevis* has retained at least 56.4% of homeologous protein-coding genes which provides a unique opportunity to study the effects of allopolyploidy on homeologous gene expression and evolution (Session et al., 2016).

Homeologs in Xenopus laevis

As a result of the allotetraploid event that created *X. laevis*, homeologous L and S chromosomes contain pairs of homeologous genes that were once nearly identical between two closely related progenitor species at the time of allotetraploidization. To identify these homeologous gene pairs nearly 17MY after allotetraploidization, the protein coding genes of the nearest diploid relative, *Xenopus tropicalis,* were placed in a 1:2 correspondence with homeologous genes within *X. laevis.* This was done using a Blastp protein alignment where homeologous pairs were identified if two *X. laevis* query protein sequences aligned to a single *X. tropicalis* gene with at least 80% identity and covered at least 50% length of the *X. laevis* query. To further validate the candidate homeologs, the synteny (similarity of neighboring genes around each homeolog) were compared between candidate homeologs. In addition to identified homeolog pairs, 6,807 orthologous genes in *X. tropi1calis* that did not have a "sister" homeolog (lacking a pairing homeolog) were identified as singletons. In total, 8,806 homeolog gene pairs were identified in *X. laevis*.

This identification method of homeologs in *X. laevis* is consistent with the definitions in the field for homeologs (Berthelot et al., 2014; Bottani, Zabet, Wendel, & Veitia, 2018; Lien et

al., 2016). Homeologs are defined as genes or chromosomes in the same species that originated by speciation and were brought back together in the same genome by allopolyploidization (Glover et al., 2016). However, most of the study done on homeologs stems from studying plant allopolyploidization where the evolutionary history of many plants is littered with polyploidization events(Magadum, Banerjee, Murugan, Gangapur, & Ravikesavan, 2013). In addition to the formal definition of a homeolog, it is important to distinguish this with other terms that describe the relationship between similar genes. Commonly confused are orthologs and paralogs which describe pairs of similar genes found in different species the originated from a speciation event or duplication event, respectively.

Homeologous pairs in *X. laevis* provide a novel opportunity for evolutionary mechanisms to act upon. Pseudogenization occurs either through mutation or deletion, rendering the effected gene as unexpressed or functionless (Magadum et al., 2013). Many homeologs that are not under selection can become lost through pseudogenization which has occurred at a rate of 64% among the subgenomes of *X. laevis*. Subfunctionalization of homeologs occurs when functions of a duplicated gene are divided among each of the sister homeologs. This is observed both spatially and temporally in the *numbl* and *six6* homeologs in *X. laevis*(Session et al., 2016). Neofunctionalization is viewed as the most interesting effect of genome duplication where one of homeologs evolves novel functions such as *hoxb4* in *X. laevis* which has gained maternal expression patterns(Session et al., 2016). Finally, gene function conservation and homeolog pair maintenance due to selection for gene dosage results in both homeologs remaining unchanged throughout evolution. In *X. laevis,* high transcript expression or robustness of expression has been hypothesized as the cause of homeolog gene maintenance. These mechanisms are important forces that shape gene expression and retention following polyploidy.

Literature Review

Characterization of Homeologs in Xenopus laevis

The publication of the *X. laevis* genome (Session et al., 2016), provided a preliminary distinguished homeolog pairs, has allowed for deeper enquiry into the evolutionary consequences of homeologs and how they are regulated in a vertebrate organism. Since then, this has caused a burst of genomic and transcriptomic discoveries that have detailed the response of homeologs in various signaling pathways such as fibroblast growth factor (FGF)(Suzuki et al., 2016), transforming growth factor-ß (TGF- ß)(Suzuki et al., 2016), Wnt(Kjolby & Harland, 2017), *hox*(Kondo, Yamamoto, Takahashi, & Taira, 2017). In addition, in depth analysis of cis-regulatory regions of *six6*(Ochi, Kawaguchi, et al., 2017) and *hand1*(Ochi, Suzuki, Kawaguchi, & Ogino, 2017) homeolog pairs have led to insight in the into the mechanisms of diverged expression differences between homeologs despite high similarity of coding regions. The purpose of this section is to review these studies and how their results led to the questions that drove the current study.

Even before the publication of the *X. laevis* genome, researchers have sought to investigate homeologs despite the obstacles of manually identifying homeolog pairs. In 2006, Chain & Evanns, performed a molecular phylogenetic analysis that asked if theorized mechanisms affecting genome duplication such as neofunctionalization or conservation through diversifying selection can be seen in the coding regions of a set of 290 retained homeolog pairs. They discovered that individual homeologs appeared to be under distinct evolutionary constraints which asymmetric rates of evolution(Chain & Evans, 2006). However, they did not analyze the cis-regulatory regions of the homeolog pairs.

In 2015, Nakede *et al.* provided evidence of homeolog-specific targeted mutagenesis using transcription activator-like effector nucleases (TALENs)(Nakade et al., 2015). TALENs are able to induce targeted double strand breaks into DNA by using a fused TAL effector DNA-binding domain and nuclease(Joung & Sander, 2013). While this study was mainly a demonstration of the potential of TALENs to independently modify a pair of homeologs with high sequence similarity, it opened the possibilities to examine the individual functions of each homeolog. These studies laid the groundwork for examining the relationships between homeologs, but were still lacking a sufficient reference genome and expression data to ask any functional and regulatory questions.

Homeolog Gene Expression Bias

Starting in 2016, genetic research on *X. laevis* was fueled by not only the completed genome assembly, but also transcriptomic data from RNA-sequencing (RNA-Seq) following the 45K known genes and ~8K homeolog pairs throughout development.

Suzuki *et al.* analyzed the TGF- ß and FGF signaling pathways in *X. laevis* which notably play role in embryonic development, patterning and cell proliferation/differentiation(Guo & Wang, 2009; Harland & Grainger, 2011). Transcriptomic analysis of homeolog pairs identified in the extracellular regulatory factors of the TGF-ß pathway showed differential expression or singleton status in 21/37 (57%) homeolog pairs. This was contrasted with results from homeolog pairs identified in the ligands of the TGF-ß and FGF families which showed a lower proportion of differential expression 11/32 (34%) and 5/20 (25%), respectively. This study highlighted the differential response of homeolog pairs in important developmental pathways to

allotetraploidization in *X. laevis* and suggest different selection pressures that effect specific parts of these pathways(Suzuki et al., 2016).

Kondo *et al*. provided the first comprehensive analysis of *hox* genes in a vertebrate throughout embryonic development in *X. laevis.* Because of the allotetraploidization event, the number of *hox* gene is *X. laevis* has doubled from 38 (*X. tropicalis)* to 76 total genes (38 homeolog pairs). Differential expression between homeologs was observed in 16/38 (42%) homeolog pairs at some point during development. Notably, clustering analysis of expression profiles over development revealed that only 10/29 (34%) (9 genes not analyzed due to low expression) homeolog pairs clustered in the same group. Taken together, both the level of expression and correlation of expression between homeologs is not conserved in many homeologs in a widely conserved developmental pathway.

Interestingly both Suzuki *et al.* and Kondo *et al.* both speculate that from their findings that their observations of differential expression between homeolog pairs is indicative of pseudogenization of one of the homeologs that will eventually lead to gene loss. They hypothesize that the mechanism of expression differences lies in the cis-regulatory regions of the homeologs where mutation accumulation is responsible for the differences between homeologs. Using homeolog specific *in* situ hybridization, Kondo *et al.* gave evidence for pseudogenization for *hoxb5.L* and *hoxb5.S* by showing the spatial localization of the two homeologs was nearly identical while confirming differential expression observed in the RNA-Seq data. Kondo *et al.* also hypothesized that the observed homeolog differential expression is due to subfunctionalization. Subfunctionalization is proposed because no gene loss was observed between homeolog pairs analyzed which may indicate that expression differences to be a coping mechanism for gene dosage as a result of having double the amount of *hox* genes in *X. laevis* as

compared to the diploid ancestor *X. tropicalis*(Kondo et al., 2017). These discussions on mechanisms for homeolog expression differences has highlighted the importance of future studies to start looking at the differences between cis-regulatory regions between homeologs and what their contribution is to expression differences.

Homeolog Cis-Regulatory Regulation

Identification of differential homeolog expression provided insights into the fate of homeologs after allotetraploidization and how this effect differs for different pathways and the components within these pathways. However, mechanisms describing the differences in homeolog expression in magnitude, temporally, and spatially were not identified by these studies. Moreover, authors of these studies speculated that these differences would be further elucidated by comparing cis-regulatory regions which is the topic of this section.

Ledford *et al.* examined cis-regulatory regions in the *six6* gene which has a distinct role during eye formation conserved across many vertebrate species. In *X. laevis*, 2 regulatory regions in the 5' upstream region and 1 regulatory in the 3' downstream were identified and conserved between multiple vertebrate species. Using transgenic constructs, it was found that gene expression was under modular control that was specific to the different gene regulatory regions identified. However, this used the sequence from *X. tropicalis* due to differences in the L and S homeolog upstream regions and so homeolog differences were not analyzed. Although a homeolog specific knockout of the downstream regulatory region using a CRISPR/Cas9 system in *X. laevis* showed a larger decrease in reporter expression biased towards the L homeolog. This result, along with the observation of expression bias towards the L homeolog during normal embryonic development and expression bias towards the L homeolog in transgene expression

constructs made from the 5' conserved regions(Session et al., 2016), suggest the importance of cis-regulatory regions in gene expression despite very similar coding regions(Ledford et al., 2017).

On the heels on Ledford *et al.,* Ochi *et al.*, delved further into the relationship between cis-regulatory regions in *six6* and expression in *X. laevis* homeologs in addition to how this may correlate with coding mutations. Expanding upon previous results, Ochi *et al.*, demonstrated endogenous expression bias towards six6.L that was four times greater that six6.S expression in whole mount in situ hybridization (WISH). This result was supported by attenuations in enhancer regions that were previously characterized(Ledford et al., 2017) but showed here to be associated with conserved transcription factor-binding motifs. Importantly, the expression of these homeologs did not differ spatially or temporally, which does not suggest that subfunctionalization or neofunctionalization plays a role the observed differences. Most notably, results demonstrating that six6.L is more important for eye growth as compared to six6.S by knockout of endogenous genes and overexpression was not only supported by enhancer differences, but also hypomorphic mutations that were found in six6.S that induce a frameshift. Overall, this study highlights the major mechanisms for difference in homeolog expression differences following allotetraploidy in addition to bringing to light the relationship between cis-regulatory and coding mutations.

Around the same time of the six6 publications, Ochi *et al.* also focused on the *hand1* homeologs which are transcriptions factors that exactly share a DNA-binding domain and suggested to be important in heart development and formation(Breckenridge et al., 2009). Expression of Hand1.L is biased towards L which also has specific spatial localization in the heart as compared to S in developing and adult *X. laevis* frogs. By first identifying the enriched

transcription factors in the enhancer regions, it was found that in the 6 identified, 1 named Myod1 had a single substitution in hand1.S as compared to hand1.L. Using a transgenic reporter containing this region from the transcription start site, expression was biased towards hand1.L reporter and most importantly only hand1.L reporter was localized to the heart. Results also showed that a rescue of the single substitution to the hand1.S reporter has spatially similar expression as hand1.L, however support for this result was poor. This study provided evidence for the importance of small changes in cis-regulatory regions that are the mechanisms behind the beginning of homeolog divergence both in expression levels and spatial patterns.

Taken together, the current literature on the specific elements that play a role in homeolog expression differences has found that cis-regulatory changes are highly important in explaining these expression differences. These studies focused on only two genes, where the S homeologs are expressed significantly lower than the L homeologs. When this observation is paired with similar expression patterns observed between homeologs, this suggest an ongoing process of pseudogenization in the S homeologs. Even though these studies do not focus on other aspects of evolution following genome duplication such as subfunctionalization or neofunctionalization, they give insights into how purifying selection is acting upon duplicated genes. Interestingly, analysis was confined to differences in coding within cis-regulatory regions. While these likely have the largest effect, epigenetic differences have yet to be
fully examined between homeologs which are known to have a significant effect on gene expression(Jaenisch & Bird, 2003).

Homeolog Expression Variability

Watanabe *et al*. performed a large scale analysis on 412 transcription factors (TF) grouped into 14 families on the basis of their DNA-binding domain in *X. laevis*. Utilizing RNA-Seq data from two clutches (biological replicates) from Session *et al*., correlation between homeolog expression patterns during embryonic development and differences in expression pattern were measured among different TF families. Comparison of homeolog TF expression showed a high correlation of expression pattern over development and similar expression levels which was also found to be significantly higher in TFs when compared with all homeologs. This conservation suggest that these genes were still under selective pressure that could be due to dosage compensation as the amount of regulatory sequence doubled due to allotetraploidization. Notably, variation of homeolog expression between the 2 clutches was observed in 26% of all homeolog TFs compared to 52% in all homeologs during development(Watanabe et al., 2016). However, there was no comparison of these variation levels to singletons or non-homeolog genes.

Following the above publication, with many of the same authors Michiue *et al*. then sought to use the same RNA-Seq data to characterize homeolog expression in cell signaling pathways, as preliminary data form the previous study on homeologs by Michiue *et al.* suggested high variability in these pathways in comparison to TFs. Analysis focused on the signaling components of the Notch, Wnt, Hippo, and Hedgehog pathways that totaled 213 homeolog pairs and 29 singletons. L homeologs in these pathways were retained at a higher rate than the S homeologs, and this proportion was overall higher than compared to all *X. laevis* homeologs. Comparing expression levels during embryonic development between homeologs revealed the majority of homeologs have correlated expression profiles and differential expression. This is

similar to the expression patterns when analyzing all homeologs, however significantly different from homeolog TFs that show a high rate of expression profile correlation as well as similar expression levels. This suggest that most homeolog components of cell signaling pathways are under pseudogenization or subfunctionalization due to differences in expression levels between homeologs(Michiue et al., 2017).

Michiue *et al*. also performed preliminary analysis of epigenetic mechanisms that might drive expression level differences resulted in observations of DNA methylation in the promoter specific to a homeolog where expression is silenced. Notably, in *dlc.L* and *dlc.*S, epigenetic differences in promoter enrichment were indeed associated with expression differences despite conservation of enhancer sequences. This suggest that differences in homeolog expression might start with epigenetic differences that could then lead to sequence differences in cis-regulatory regions(Michiue et al., 2017).

Interestingly, it was noted that only 56% of signaling pathway homeolog expression patterns and levels were consistent between the two biological replicates used for the analysis. This is a similar compared to the 48% rate observed among all homeologs in *X. laevis*, but much lower than the 74% consistency observed in homeologs TFs found in the previously reviewed study. This highlights not only the variability between expression profiles of homeologs in relation to each other, but the variability homeolog expression between independently developing embryos. As the focus of this study was to characterize the differences and variability in homeologs apart of developmental signaling pathways, it remains unclear as to how gene expression variability in homeologs can compare to non-homeologs and singletons. It also raises further questions about the potential mechanisms that not only drive homeolog expression differences, but also high rates of variability(Michiue et al., 2017).

Literature Review Summary

The studies mentioned in this section, which is to my knowledge the extent of homeolog research in *X. laevis* since publication of the genome, provide a basis on which to form other questions regarding homeolog gene regulation. The studies of homeolog expression in specific pathways characterized the differences of homeolog expression within important developmental pathways while the studies of cis-regulatory regions of homeologs gave evidence that suggested what may be causing these differences. Interestingly, Ochi *et al.* makes connections between mutations found between homeologs in cis-regulatory regions studies and human disease phenotypes that may be explained by these cis-regulatory elements rather than the traditional focus on coding mutations. Further relationships found between dosage differences in homeologs and the mutations accumulated in homeolog coding and non-coding and human disease variants may provide important insight into how genes maintain proper dosages and which mutations are important(Lever & Sheer, 2010).

While this research has characterized much of how homeologs behave in relation to each other during embryonic development, many questions are left unanswered. This includes the further characterization of homeolog variability that is observed between biological replicates, how does this compare to variability in non-homeologs, and what mechanisms drive this variation? Initial observations of variable expression from an evolutionary perspective might suggest competition among homeologs for dominance, whereas a developmental perspective might see variability as a mechanism to maintain a buffer against genetic and environmental perturbations. In regards to developmental perturbations, nothing is known homeolog behavior in response to developmental perturbation and what importance they may provide in compensation. Where some cis-regulatory regions were looked at in detail for single genes, it is unclear whether

these findings can be generalized among most homeologs. This requires a large-scale analysis of cis-regulatory regions between homeologs where expression levels and patterns have been previously characterized. Providing explanations to these important questions will expand our knowledge of the consequences of allotetraploidization in a vertebrate organism, particularly its significance during embryonic development.

Goals and overview

The above findings of gene expression relationships between homeologs demonstrates that following allotetraploidization, asymmetric expression levels and patterns depend on the class of homeologs. Homeologs encoding transcription factors, with the exception of hox genes, maintain a high rate of conservation in homeolog expression patterns and levels, while homeologs in signaling pathways highly important in development largely exhibit decreased expression of the S homeolog and highly variable expression patterns. The latter may be explained by mechanisms of subfunctionalization and pseudogenization which in turn are driven by changes in cis-regulatory enhancers of repressors that control gene expression.

There were 2 main aims that this project sought to fulfill: 1) Comprehensive characterization of homeolog expression across different gene classes has provided basic expression differences between homeologs, however further characterization of highly variable homeolog expression during embryonic development and the mechanisms underlying these patterns remain to be elucidated. 2) Characterization of homeolog regulation in response to developmental perturbation is an untouched area of research that will provide further insight into the role of homeologs in compensation and whether their diversity contributes to robustness.

Further characterization of highly variable homeolog expression was prompted by the strikingly low rate of consistency (56%) of expression patterns and levels of homeologs between biological replicates during embryonic development as first presented in Session *et al*. 2016, and highlighted in Michiue *et al*. 2017. As only 2 biological replicates were used for the first analyses that described high variability, additional replicates were retrieved from previous RNA-Seq experiments uploaded to the Sequence Read Archive (SRA) hosted by the National Center for

Biotechnology Information (NCBI). This provided a more robust dataset in which to characterize gene expression variability.

Previous unpublished work in the Saha Lab has investigated the time course of response to genetic and physical perturbations to embryonic development in *X. laevis* using RNA-Seq. In the genetic study, the genetic perturbation focused on hyperactivation and downregulation of the Notch signaling pathway where embryos exhibit compensation from both types of perturbation. Briefly, Notch signaling is an evolutionary conserved juxtacrine signaling pathway which controls differentiation and proliferation of cells in multiple tissues during embryonic development(Andersson, Sandberg, & Lendahl, 2011). This pathway is especially important in the context of neurogenesis where notch signaling is required for maintenance of neural stem cell populations and neural commitment inhibition that both work to maintain the balance of proliferative and differentiated neural cells(Lasky & Wu, 2005).

In the study on physical perturbations, the response to a 180-degree rotation of the neural anterior-posterior (AP) axis during gastrulation was investigated where a specific time frame of plasticity allows compensation from this. Briefly, neural patterning along the AP axis provides identities to pluripotent cells based on their position along the neural tube. This is a crucial process during embryonic development, specifically neural induction, which divides the neural tube into hallmark architecture in the central nervous system such as the forebrain, midbrain, hindbrain, and spinal cord(Altmann & Brivanlou, 2001)(Hendrickx, Van, & Leyns, 2009).

The two above studies employed global RNA-Seq at various time points following initial perturbation which has allowed comprehensive profiling of the transcriptome. In addition to characterizing the response to genetic perturbation close to the initial perturbation, this has identified the genetic programs that are involved in the compensatory response. However, until

the recent publication of the genome, analysis has been limited to treating homeolog pairs as single genes. In this project, characterization of homeolog regulation in response to developmental perturbation was made possible by using the newly assembled genome(Session et al., 2016). With this, homeolog expression bias was identified by comparing the homeolog expression levels directly against each other and change in homeolog expression bias was identified by comparing control and experimental conditions. A temporal aspect was introduced by utilizing the time points following initial genetic perturbation where homeolog bias or change in homeolog bias can be observed during the compensatory period which has lacked study in similar perturbation experiments(Riddiford & Schlosser, 2017).

Materials and Methods

Homeolog Gene identification

A curated version of the *X. laevis* v9.1 genome assembly with gene models JGIv18pV3 (accessed 170527) (45,829 transcripts) (Atsushi Suzuki, Masanori Taira, Taejoon Kwon) of the JGlv 1.8.3.2 annotation (Xenbase) was used to identify genes. Homeologs were identified using an initial list of 8,806 putative homeologs and 6,807 singletons were obtained from work done by Session *et al*. 2016 in the publication of the Xenopus laevis genome. This list was expanded using a custom python script (supplement) which searched the JGIv18pV3 annotation for genes with the same symbol but on opposite chromosomes where the ".L" suffix was required for the gene on the L chromosome and the ".S" suffix was required for the gene on the S chromosome. This method for homeolog identification differs from the method used in the initial homeolog identification (Session et al., 2016). However, upon examination of newly discovered homeolog gene models, these fit the same criteria of synteny and alignment rate used in the initial homeolog identification (Figure 1).

RNA-Seq data analysis - Notch Experiment

Library Prep and high-throughput sequencing

Samples were collected from 3 matings, where replicate number represented the respective mating. 5 embryos representing each sample were then pooled and RNA was extracted using x kit. Paired-end 50bp sequencing using the deoxy UTP (dUTP) strand-marking protocol (Parkhomchuk et al., 2009) was performed using either 3 or 4 lanes of a Hiseq 2000 (Illumina) yielding an average library depth of ~56 million reads per sample.

Read processing, mapping & quantification

Quality assessment of RNA-Seq reads was performed using FastQC (version 0.11.5) (Andrews, 2010). Each individual pair was assessed for per base sequence quality and adaptor content. All reads had phred quality scores greater than 28; no trimming was done.
Paired-end mapping was performed using Hisat2 (version 2.0.5) (Kim, Langmead, & Salzberg, 2015) using default settings against the *Xenopus laevis* v9.1 genome downloaded from Xenbase, yielding an average of 95.4% alignment rate and 47.91% average coverage. Output sequence alignment map (SAM) files were converted to binary alignment map (BAM) format and then sorted by name using samtools (version 1.5) (Li et al., 2009).

Read quantification for non-collapsed transcripts was performed using HTSeq-Count (version 0.6.0) using the JGIv18pV3 (accessed 170527) annotation (45,829 transcripts) which is a manually curated version (Atsushi Suzuki, Masanori Taira, Taejoon Kwon) of the JGlv 1.8.3.2 annotation (Xenbase). HTSeq-count was used to count reads aligning to exon regions using parameters '-f bam' for BAM file input, '-s reverse' for handling paired-end reverse stranded reads, and '-I name' for using the GFF name attribute to identify counts (Anders, Pyl, & Huber, 2015) .

Differential gene expression analysis

Differentially expression of control (GFP) vs experimental conditions (ICD, DBM) and conditions across stages were tested using un-normalized counts from HTSeq-count using DESeq2 (version 1.16.1) (Love, Wolfgang, & Andrers, 2014). As the 3 stages differed in normalization distribution and dispersion distribution (Supplementary normalization fig), samples were loaded into DESeq2 by stage and comparisons were made within stage, with the

exception of across stage comparisons of the same condition where only 2 conditions of interest were loaded per comparison (Supplementary across stage).

A general linear model with the design '~ replicate + condition' was used to test for effects between conditions while controlling for the replicate number that was associated with the clutch the embryo originated from. After filtering genes, whose sum of counts over the replicates per gene was less than or equal to 10, the 3 biological replicates in each of the 9 conditions allowed for statistically relevant pairwise comparisons using the wald test. Resulting features were significantly differentially expressed if the benjamini-hochberg adjusted p-values (padj) were less than the parameter 'alpha = 0.05' using the 'results()' function. Except for the GFP vs ICD stage 18 comparison and across stage comparisons, p-value distributions of raw p-values after differential expression testing were right skewed, suggesting an overestimation of dispersion values for some genes. This was corrected for using the default settings of fdrtool with z-scores calculated by DESeq2 as input to estimate the empirical null distribution and recalculate p-values (version 1.2.15) (Strimmer, 2008).

RNA-Seq data analysis – Anterior-Posterior Rotations Experiment

Library Prep and high-throughput sequencing

77 samples were collected in two batches from x mattings, where replicate number was not representative of the respective mating. RNA was extracted from single embryos using x kit. Paired-end 75bp sequencing using the deoxy UTP (dUTP) strand-marking protocol (Parkhomchuk et al., 2009) was performed using either 3 or 4 lanes of a Hiseq 2000 (Illumina) yielding an average library depth of ~64 million reads per sample for the first batch containing 50 samples. It should be noted that in the first batch 7/100 fastq files were corrupt. About half of

the data was able to be recovered using gunzip (version 1.7) recovery protocol. Paired-end 150bp sequencing using the deoxy UTP (dUTP) strand-marking protocol (Parkhomchuk et al., 2009) was performed using either 3 or 4 lanes of a Hiseq 2000 (Illumina) yielding an average library depth of ~80 million reads per sample for the second batch containing 22 samples.

Read processing, mapping & quantification

Quality assessment of RNA-Seq reads was performed using FastQC (version 0.11.5) (Andrews, 2010). Each individual pair was assessed for per base sequence quality and adaptor content. All reads had phred quality scores greater than 28; no trimming was done.

Paired-end mapping was performed using Hisat2 (version 2.0.5) (Kim, Langmead, & Salzberg, 2015) using default settings against the *Xenopus laevis* v9.1 genome downloaded from Xenbase, yielding an average of 93.9%% alignment rate and 44.8% average coverage. Output sequence alignment map (SAM) files were converted to binary alignment map (BAM) format and then sorted by name using samtools (version 1.5) (Li et al., 2009). It should be noted that in the 7 recovered fastq files, many paired-end read mates were lost for the affected samples. These were separated using samtools (version 1.5) (Li et al., 2009) before counting as HTSeq-Count is unable to count combined paired-end and single-end reads.

Read quantification for non-collapsed transcripts was performed using HTSeq-Count (version 0.6.0) using the JGIv18pV3 (accessed 170527) annotation (45,829 transcripts) which is a manually curated version (Atsushi Suzuki, Masanori Taira, Taejoon Kwon) of the JGlv 1.8.3.2 annotation (Xenbase). HTSeq-count was used to count reads aligning to exon regions using parameters '-f bam' for BAM file input, '-s reverse' for handling paired-end or single-end reverse stranded reads, and '-I name' for using the GFF name attribute to identify counts (Anders, Pyl, & Huber, 2015). For 7 samples where single-end and paired-end reads were

analyzed separately, single-end reads were added to paired end reads to obtain final raw read counts.

Differential gene expression analysis

Differentially expression between conditions (see table x for all conditions) across stages 18 and 30 were tested using un-normalized counts from HTSeq-count using DESeq2 (version 1.16.1) (Love, Wolfgang, & Andrers, 2014). Preliminary data exploration and clustering showed strong batch effects between the first and second batches. Sibling conditions were shared between these two batches (5 in the first batch and 2 in the second batch), which were leveraged to remove mean shifts associated with the categorical "batch" covariate. A likelihood ratio test with the design '~ batch + condition' was used to test for effects between conditions while controlling for the batch effects in test involving samples from both batches. Otherwise the design '~ condition' was used to test for effects between conditions within the same batch. After filtering genes, whose average counts over any 2 of the replicates per gene was less than 5, the 5 biological replicates in each of the 14 conditions allowed for statistically relevant pairwise comparisons where resulting features were significantly differentially expressed if the benjamini-hochberg adjusted p-values (padj) were less than the parameter 'alpha = 0.05' in the 'results' function.

Homeolog Differential Expression testing

Raw read counts were obtained for each experiment using the same read processing, mapping, and quantification methods described in previous sections. Homeologous gene pairs were extracted from read count list and placed into their own respective list using a custom python script. Because differential expression analysis usually assumes that the same gene will

be compared across different conditions, read length is usually not accounted for (Love, Wolfgang, & Andrers, 2014). However, to test differential expression between homeologs where transcript lengths are different between homeologs, transcript lengths specific to each homeolog must be used in read count normalization to account for lengths that influence read counts. Instead of using the "median ratio method" employed by DESeq2 to normalize read counts between samples and calculate a size factor per sample, normalization factors were calculated per gene using gene lengths for each homeologous gene. After loading separate L homeolog and S homeolog subgenome samples representing a single "whole" sample into DESeq2, genes were then filtered if the sum of counts over the replicates per gene was less than or equal to 10. For the Notch experiment, a general linear model with the design '~replicate + condition + condition:replicate + condition:subgenome'. The 'replicate' term was to account for replicates that represent different clutches among conditions (identified using PCA), the 'condition' term was included to test for differences across conditions, the 'condition:replicate' interaction term is to account for variance among the samples in the control and experimental group, and the 'condition:subgenome' interaction term is to estimate the difference in homeolog bias ratios across different conditions. Resulting homeologs were significantly biased or differentially biased if the benjamini-hochberg adjusted p-values (padj) were less than the parameter 'alpha = 0.05' using the 'results()' function. Except for the homeolog bias comparisons, p-value distributions of raw p-values after differential expression testing were right skewed, suggesting an overestimation of dispersion values for some genes when fitting the counts to negative binomial model. This was corrected for using the default settings of fdrtool (version 1.2.15) (Strimmer, 2008) using z-scores calculated by DESeq2 as input to estimate the empirical null distribution and recalculate p-values.

Homeolog Differential Variance Testing

Raw read counts were obtained for each experiment using the same read processing, mapping, and quantification methods described in previous sections. Homeologous gene pairs were extracted from read count lists and placed into their own respective list using a custom python script. Read counts between samples were normalized using the "median of ratios" method (Love et al., 2014) using DESeq2 and then normalized for gene length. MDSeq (Ran & Daye, 2017) was used with the design '~subgenome' for differential variability testing between L and S homeologs using pre-normalized counts as described above. Resulting homeologs had significant differential variability if the benjamini-hochberg adjusted p-values (padj) were less 0.05.

Transcript Annotation and Functional Enrichment

The transcripts from the *X. leavis* genome v9.1 were functionally annotated using the Blast2GO CLC plugin (version 1.10.6) with default parameters. Briefly, 45,107 cDNA sequences were blasted using blastx with E-value cutoff 1.0E-5 against the NCBI database where 45,092 sequences returned hits. 30,137 gene ontology (GO) terms were then mapped to blastx hits and 27,484 annotated using default parameters.

Two-tailed fisher's exact [Office2] test within Blast2GO was used to test for enrichment of significant differentially expressed (SDE) genes (padj < 0.05) resulting from pairwise comparisons. Test sets were resulting SDE genes from a comparison and reference sets included genes with average expression greater than 10 over all samples in the respective stage. This reduced the reference to only genes that were expressed during the experiment, giving more meaningful enrichment. Significantly enriched GO terms and false discovery rate values of

enriched GO terms were visualized using REVIGO (accessed 171221) using parameters '0.7' for allowed similarity, 'SimRel' for semantic similarity clustering, and 'whole UniProt' as the database with GO term sizes (Supek, Bošnjak, Škunca, & Šmuc, 2011).

Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016) pathway and module enrichment was done using clusterProfiler (version 3.4.4) (Yu, Wang, Han, & He, 2012) with the *X. Laevis* KEGG database (version 3.2.3). KOBAS (version 3.0) (Xie et al., 2011) was used for KEGG ID annotation that resulted in 29232/45107 KEGG IDs mapping to *X. Laevis* protein sequences. Two methods were used to test for pathway enrichment, over-representation and gene set enrichment analysis (GSEA). In over-representation analysis, reference sets identical to those used in GO analysis were used and gene sets were either thresholded at padj < 0.05 or < 0.1 before testing for enrichment. In GSEA, gene sets were thresholded by padj < 0.05, 0.1, or 0.5 and then ranked in descending order by $\log_2$ fold change. Pathway enrichment was visualized using pathView (version 1.16.7) (Luo & Brouwer, 2013).

Homeolog Variance Analysis

Read processing, mapping & quantification

Raw RNA-Seq data from control samples at mid-gastrula (5 samples), mid-neurula (1 sample), and late tailbud (1 sample) stages (Session et al., 2016; Ding et al., 2017; Peshkin et al., 2015) was downloaded from the Sequence Read Archive database and pooled with 'sibling' samples from mid-neurula (7 samples) and late tailbud (7 samples) stages in the Anterior-Posterior Rotations experiments (Bolkhovitinov, 2017). Read mapping and quantification was performed as described above.

Normalization & Batch Correction

Sequencing depth and RNA composition was normalized between all 21 samples using the DESeq2 (Love et al., 2014) median of ratios method. Preliminary clustering analysis of samples within each stage revealed experiment specific batch effects. Batch correction was performed using normalized log transformed counts with the removeBatchEffects() function in the limma package (Ritchie et al., 2015) between samples from different experiments within each stage. Batch corrected counts were then analyzed using PCA to ensure samples from different experiments now clustered together. Lastly, counts were transformed back from log space and normalized for transcript length.

Variation Analysis

Samples were processed by stage where lowly expressed genes were filtered if the expression of the gene was below 5 in 2 or more of the samples within each stage. Variation of gene expression was calculated per gene between replicate samples within the same stage. Due to the inherent nature of discrete count data, it has been empirically shown (Bar & Schifano, 2018; Love et al., 2014; Ran & Daye, 2017; Ritchie et al., 2015; Wu, Wang, & Wu, 2013; Zhou, Lindsay, & Robinson, 2014) that there is a consistent relationship between the variance and the mean which positively increases as the average count approaches 0. This relationship, where low expression means exhibit proportionally more variance than higher expression means, is referred to as heteroscedascity. Many differential expression analysis packages attempt to account for this by either fitting data to a negative binomial model that includes a parameter to account for non-Poisson dispersion (Love et al., 2014) or by estimating the mean-variance trend to produce inverse variant weights to be used in linear modeling (Ritchie et al., 2015).

Since we are focused on comparing variances of genes while accounting for gene

length, we also need to account for the mean-variance relationship to account for any patterns or differences that are simply due to a change in the mean expression. We first attempted to measure variation of gene expression using the coefficient of variance which is calculated by dividing the standard deviation by the mean. This allows the variance at different mean expression levels to be compared and is dependent on expression levels between replicate to be normally distributed. We observed in all cases a heteroscedastic relationship between the log expression mean and the coefficient of variation which made this measure unsuitable for comparing expression variance.

We use the voom transformation within the limma (Ritchie et al., 2015) package in order to estimate the mean variance trend and obtain a measure of standard deviation (sigma), which we use as our measure of variability, where the mean-variance trend has been removed. The mean variance trend was assessed before and after the voom transformation to confirm the removal of variance dependence on the mean.

Statistical testing

Distributions of mean expression and coefficient of variation were found to be consistently non-normal and thus compared using the non-parametric Wilcox Rank-Sum Test in R studio (R Core Team, 2017). Correlations between variables were tested using the non-parametric Kendall's tau test which assess the concordance and discordance between paired observations.

Results

Expanded Identification of Homeolog Genes

Preliminary analysis of genes of interest showed that many genes that appear to have homeologous pairs that were not annotated in the initial genome publication (Session et al., 2016). To expand upon the initial list of 8,806 putative homeolog pairs identified by Session *et al*., a method identifying homeolog pairs by name was developed that identified an additional 1,873 homeolog pairs. This identification was validated for many homeolog pairs using the same criteria set forth by Session *et al*. and examples of newly identified homeologs pax3, ttyh1, and parvg are shown in figure 1.

A)

B)

Homeolog	Identity %	Coverage %
pax3	98	98
ttyh1	82	92
parvg	100	86

Figure 1 | Expanded homeolog identification. Gene models of pax3, ttyh, and parvg homeolog pairs found using a search based on nomenclature. A) L homeolog synteny is shown in the top half and S homeolog synteny shown in the bottom half. Synteny is consistent between homeolog pairs. B) BlastP alignment and coverage percentages between homeologs. These percentages are consistent with the criteria put forth in the initial homeolog identification by (Session *et al* 2017).

Homeolog Expression Variance Over Time-course of Embryonic Development

In order to further investigate the high rates of homeolog expression variance previously identified (Michiue et al., 2017; Watanabe et al., 2016), additional *X. laevis* RNA-seq samples corresponding to control conditions of 3 stages during embryonic development were downloaded from the SRA database (Session et al., 2016; Ding et al., 2017; Peshkin et al., 2015) and included with 'sibling' samples from the AP-Rotations experiments previously conducted in the lab. This was done to increase the sample size from 2 to 5 samples in the mid-gastrula stage and 8 samples in the mid-neurula and late tailbud stages (Nieuwkoop & Faber, 1994). These stages were selected as they contained the highest amount of biological replicates when combined with in-house data, increasing the robustness of the analysis. Raw RNA-seq data was processed identically for all 17 samples and batch effects were removed between samples within stages as described in the methods.

Preliminary analysis showed that the amount of expressed homeologs increased over the time course of development where 66% of expressed homeologs were shared among all the analyzed stages. 14% of homeologs were exclusively shared between the mid-neurula and late tailbud stage while under 4% were exclusively shared between the mid-gastrula stage and either the mid-neurula or late tailbud stages (Figure 2b). Despite these differences in the composition of homeologs across the stages, the percentage of homeologs which made up the total amount of genes expressed remained relatively unchanged (63% ± 1).

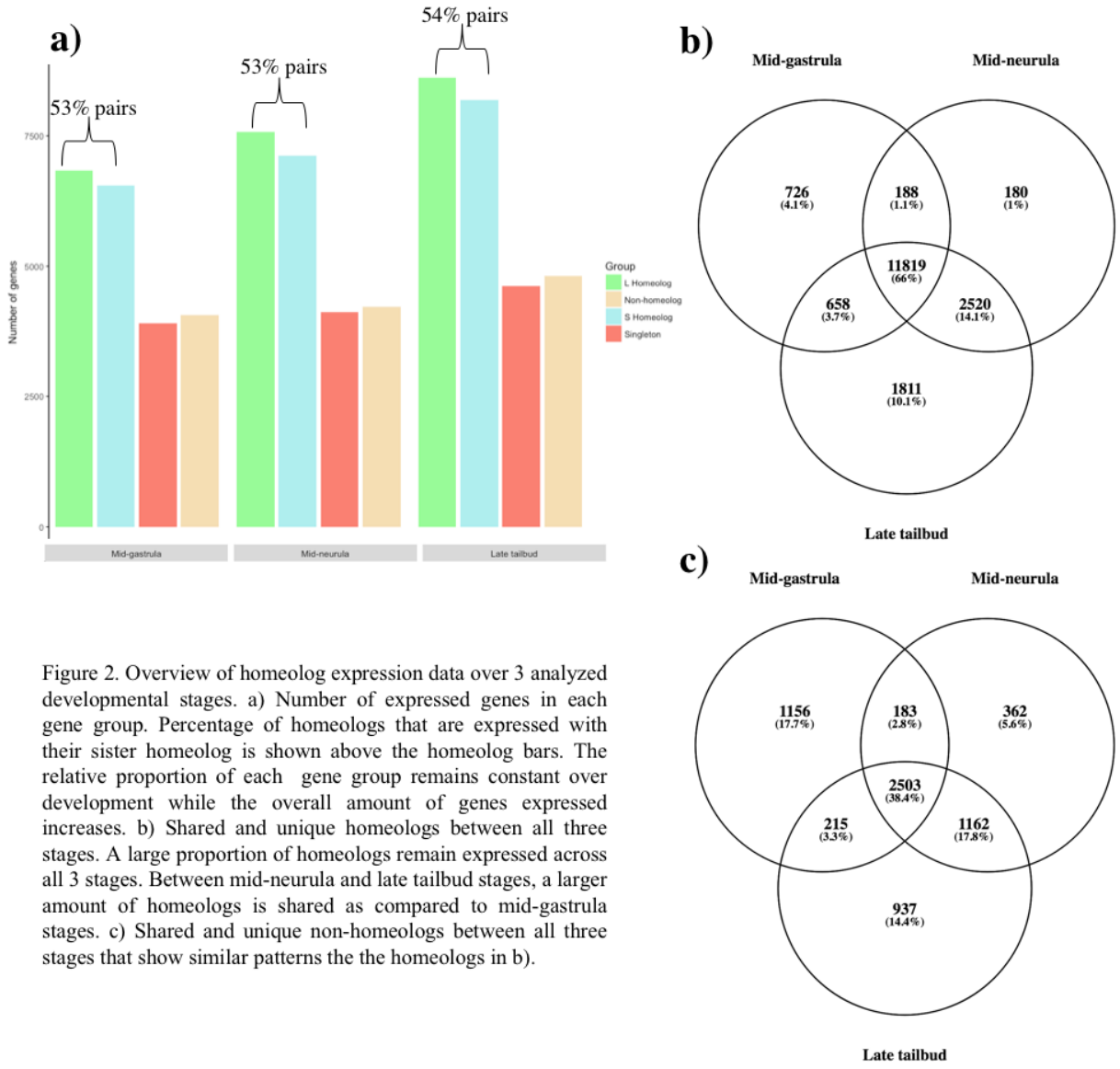


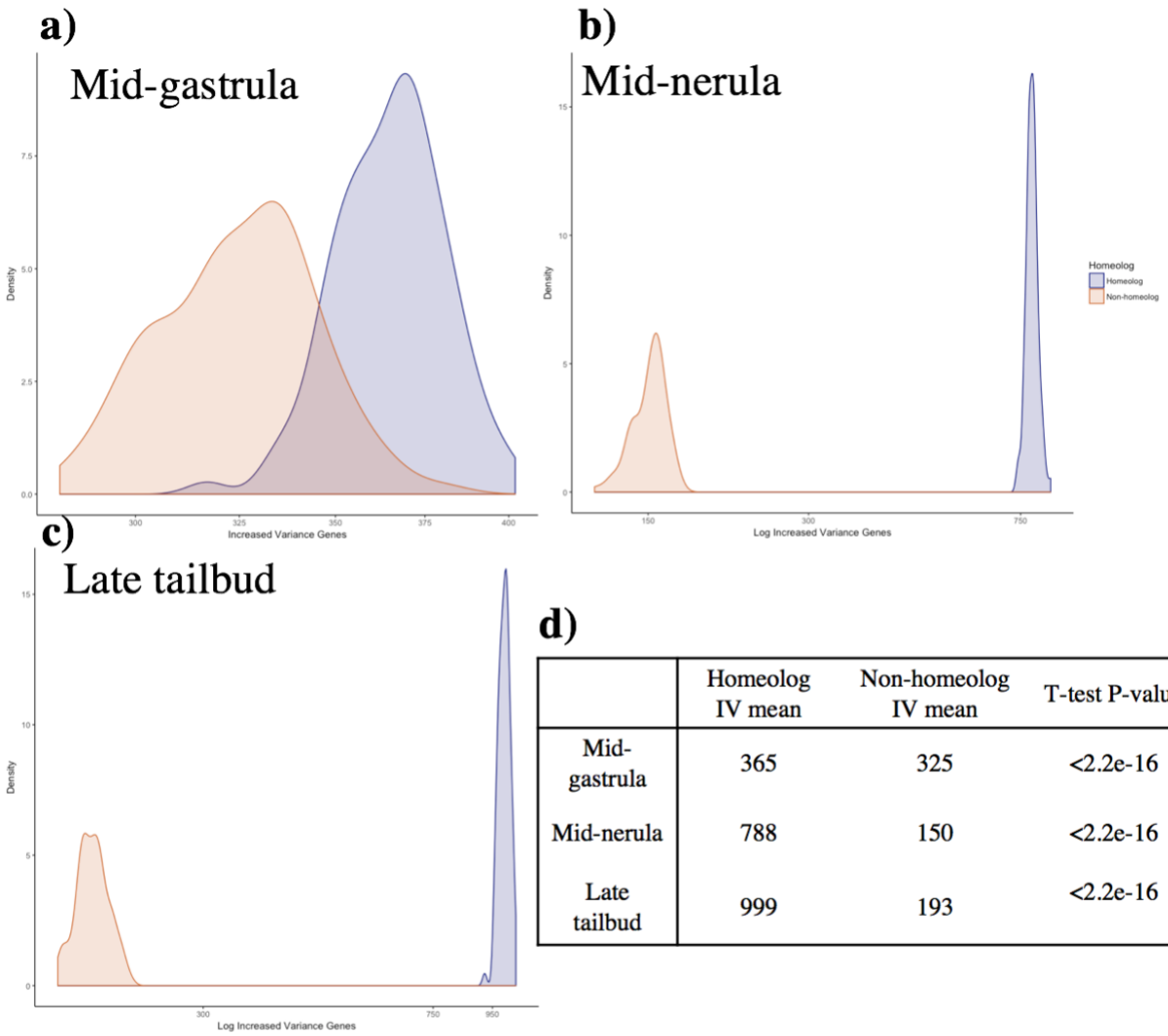
Figure 2. Overview of homeolog expression data over 3 analyzed developmental stages. a) Number of expressed genes in each gene group. Percentage of homeologs that are expressed with their sister homeolog is shown above the homeolog bars. The relative proportion of each gene group remains constant over development while the overall amount of genes expressed increases. b) Shared and unique homeologs between all three stages. A large proportion of homeologs remain expressed across all 3 stages. Between mid-neurula and late tailbud stages, a larger amount of homeologs is shared as compared to mid-gastrula stages. c) Shared and unique non-homeologs between all three stages that show similar patterns the the homeologs in b).

Homeologs Exhibit Greater Expression Variability than Non-Homeologs

As the relationship between homeologs has been found to be highly variable in previous studies (Michiue et al., 2017; Watanabe et al., 2016), we asked if this might be due to the increased variability in homeolog gene expression than would be expected. This was tested by randomly pairing homeolog genes with non-homeolog genes and testing for a difference in expression variability using MDSeq (Ran & Daye, 2017). This test was repeated 1000 times in order to obtain a null distribution of the expected increase in variation. We then measured the

amount of genes in the homeolog and non-homeolog groups that had an increase in variance each time the test was repeated. This resulted in a significantly larger amount of homeolog genes were an increase in variance was observed as compared to non-homeolog genes across all stages (Figure 3).

Mid-neurula and late tailbud stages showed on average 357 more homeologs with increased variance and 154 fewer non-homeologs with increased variance than at the mid-



d)

	Homeolog IV mean	Non-homeolog IV mean	T-test P-value
Mid-gastrula	365	325	<2.2e-16
Mid-nerula	788	150	<2.2e-16
Late tailbud	999	193	<2.2e-16

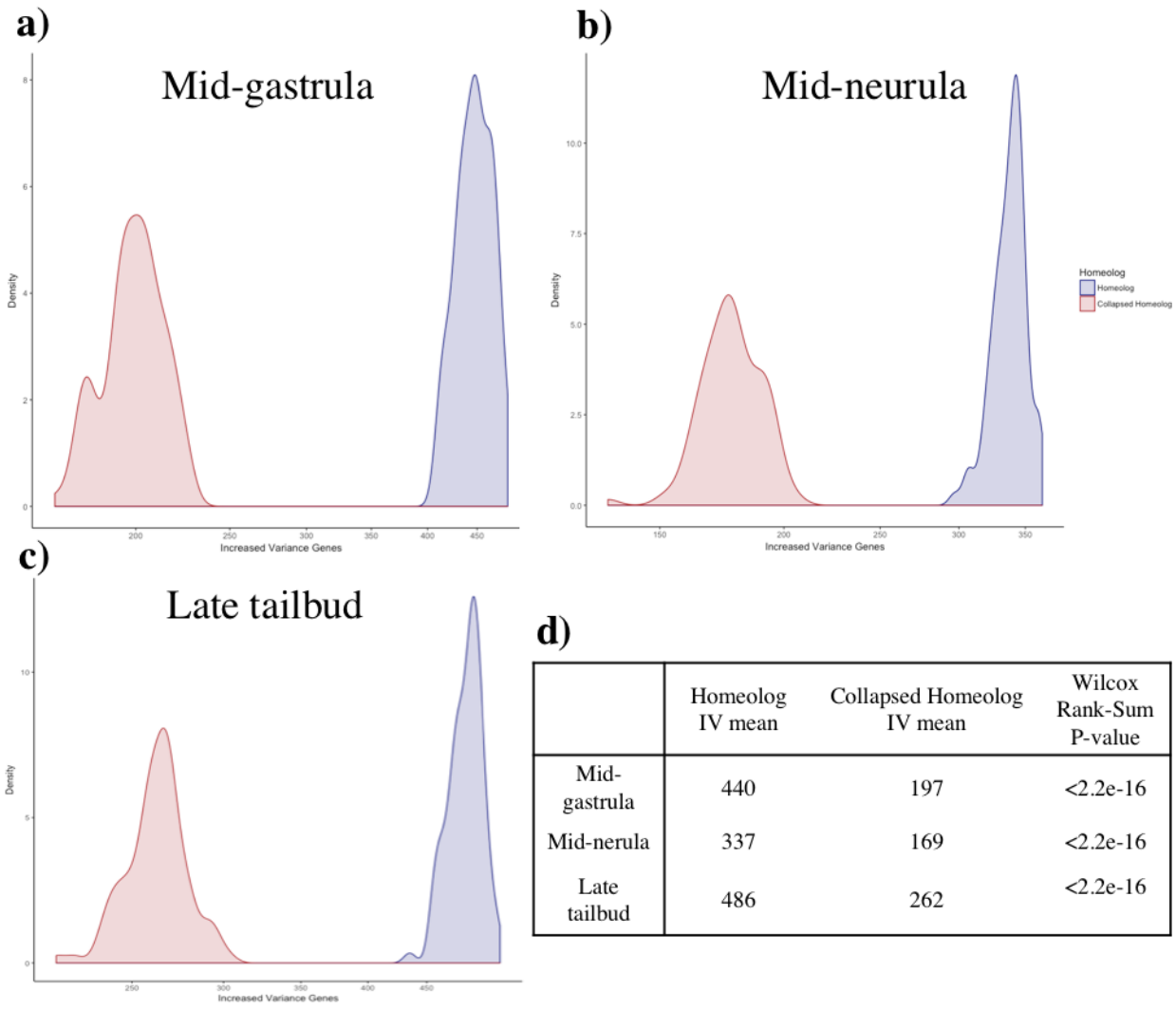
Figure 3. Homeologs Exhibit Greater Expression Variability than Non-Homeologs. Comparison of distributions of homeolog sand non-homeologs that exhibit increased variance when randomly paired and tested for changes in variance. This test was repeated 1000 times. All 3 stages are significantly different and this difference is lowest at the a) mid-gastrula stage. This increases as the mid-neurula stage and peaks at the late tailbud stage. Table d) showing means of each distribution and P-values resulting from comparison of distributions.

gastrula stage (Figure 3b, c). The large differences between these stages may be accounted for by the differing amounts of compositions of homeologs (Figure 2b) and non-homeologs across all 3 stages. While over half of all homeologs and non-homeologs (66% and 38.4% respectively) were shared among all 3 stages, a large proportion of these where exclusively shared between the mid-neurula and late tailbud stages as compared to sharing between these stages and the mid-gastrula stage. However, these differences do not change the overall trend of increased variance of homeolog gene expression.

The Sum of Homeolog Expression Exhibits Expression Stability

Because of the recent (17-18 MYA) allotetraploidy event in *X. laevis*, most homeologs that have been retained have very similar (> 80%) protein coding similarity to the nearest diploid relative orthologous genes in *X. tropicalis*. By extension, similar protein coding homeologs will have relatively similar function if one of the homeologs is not already subfunctionalized or pseudogenized. Redundant function from both homeolog gene products may be required to compensate for the doubling in the genetic workspace, e.g. the domain of a transcription factor, that requires twice as many gene products. Thus, the sum of expression values from homeolog pairs can be considered the overall expression level that would normally be produce by a single gene, i.e. a non-polyploid organism.

This sum represents the total expression of both homeologs, which can be thought of as a highly expressed non-homeolog gene. Since we have observed less variable gene expression in non-homeologs as compared to homeologs by obtaining distributions of the number of genes with increased variance, we hypothesized that the sum of expression of a homeolog pair would exhibit less variance in expression as compared to the expression variance of homeologs.



d)

	Homeolog IV mean	Collapsed Homeolog IV mean	Wilcox Rank-Sum P-value
Mid-gastrula	440	197	<2.2e-16
Mid-nerula	337	169	<2.2e-16
Late tailbud	486	262	<2.2e-16

Figure 4. The sum of homeolog pair expression is more stably expressed than individual homeolog genes. Expression levels of homeolgos were summed into 'collapsed homeologs' where were compared against randomly paired homeolog genes and repeated 1000 times to obtain a distribution of genes that showed an increase in variance either toward homoelogs or non-homeologs. Differences between homeologs and collapsed homeologs were the greatest at the mid-gastrula stage followed by the late tailbud stage and least at the mid-neurula stage b). d) Table of means for each distribution and P-values comparing distributions using a Wilcox-rank sum test.

This was tested at each stage by summing the counts replicate-wise between homeolog pairs that were both expressed, randomly pair each summed expression observation with a random homeolog gene, and test for a difference in expression variability using MDSeq (Ran & Daye, 2017).. This test was repeated 1000 times in order to obtain distributions of the number of genes with an increased variance for the homeolog group and the summed homeolog group,

referred to as ‘collapsed’ homeologs. Results as shown in figure 4 how that collapsed homeologs had significantly lower amounts of genes with increased variance after each test.

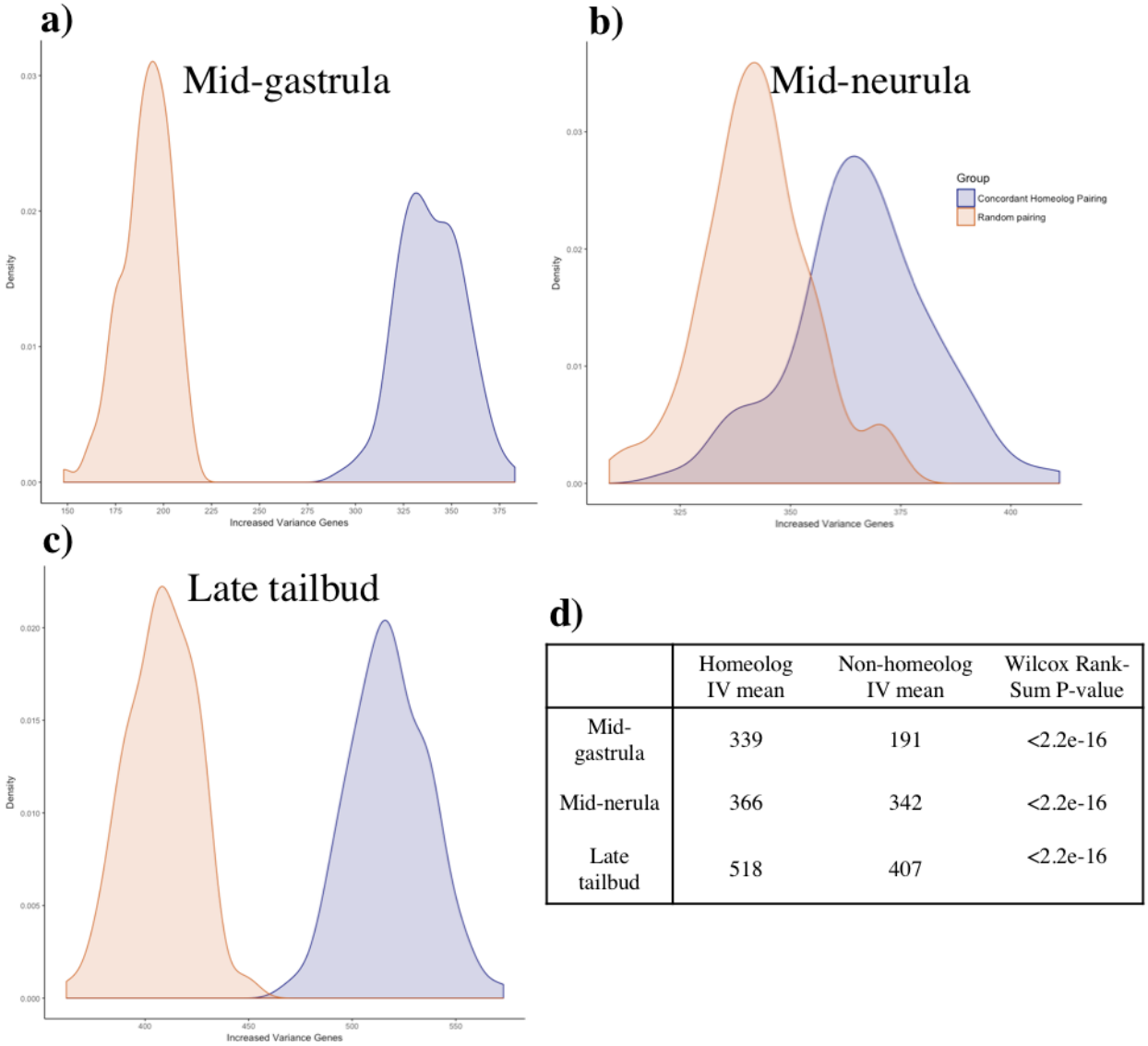
This indicates that the expression variance of collapsed homeologs is less than that of homeologs. Moreover, this suggest that the total gene product of both homeologs is expressed more stably as compared to individual homeologs even though individual homeologs are more variant than would be expected based on previous results.

The Relationship of Gene Expression Between Homeologs is Highly Variable

Variability between the relationship of homeologs has been previously characterized (Michiue et al., 2017; Watanabe et al., 2016) as inconsistent patterns of homeolog expression across the time course of development (oocyte to late tailbud) between 2 biological replicates. We sought to assess the validity of this finding by using a larger sample size. In addition, we consider our approach more robust than in previous studies as the measurement variability does not rely on mean expression and comparisons of differential variability utilizes a statistical test directly comparing variability of two genes rather than qualitatively comparing ratios of consistent expression patterns.

We define the relationship between homeologs on the same scale as gene expression that can be defined as the absolute difference in expression between a homeolog pair. A permutation test was repeated 1000 times by randomly pairing L and S homeologs, taking their absolute expression difference, and using MDSeq (Ran & Daye, 2017) to compare the variability of this absolute difference against the variability of absolute difference of concordantly paired homeologs.

Across all 3 stages, the variance of the relationship between homeologs is significantly greater than the constructed null expectation by random pairing of homeologs (Figure 5). The largest difference between the null expectation and increased variance of the homeolog relationship was at the mid-gastrula stage (Figure 5a) which was unexpected at this stage given



d)

	Homeolog IV mean	Non-homeolog IV mean	Wilcox Rank-Sum P-value
Mid-gastrula	339	191	<2.2e-16
Mid-nerula	366	342	<2.2e-16
Late tailbud	518	407	<2.2e-16

Figure 5. The relationship between homeologs is highly variable as compared to the null expectation. Increased variance genes is the number of genes which showed a significant increase when testing the the variance of the concordantly paired homeologs (blue) against randomly paired homeologs (orange). The mid-gastrula stage a) exhibited the largest difference in the number of increased variance genes followed by the late tailbud stage b) and mid-neurula c) stage. d) All differences were were significantly significant ($P < 0.05$) using the Wilcox-Rank Sum test.

the relatively small difference of increased variability as compared to non-homeologs (Figure 4a) and between the homeologs in previous results. These findings expand on the evidence that the

relationship between homeologs as measured in the difference in expression is more variable than expected.

Characterization of homeologs based on expression difference and variance of expression difference

To further characterize the highly variable relationship between homeologs observed in the previous section, we plotted the variation of the expression difference as a function of the absolute expression difference. This was then divided into 4 quartiles to examine the patterns of homeolog expression within each quartile. We observe 4 different expression behaviors within homeologs characterized as: 1) High variation of expression difference and low absolute expression difference (HVLD) (Figure 6b), 2) High expression variation difference and high expression difference (HVHD) (Figure 6c), 3) Low variation of expression difference and high expression difference (LVLD) (Figure 6d), and 4) Low variation of expression difference and low expression difference (LVHD) (Figure 6e).

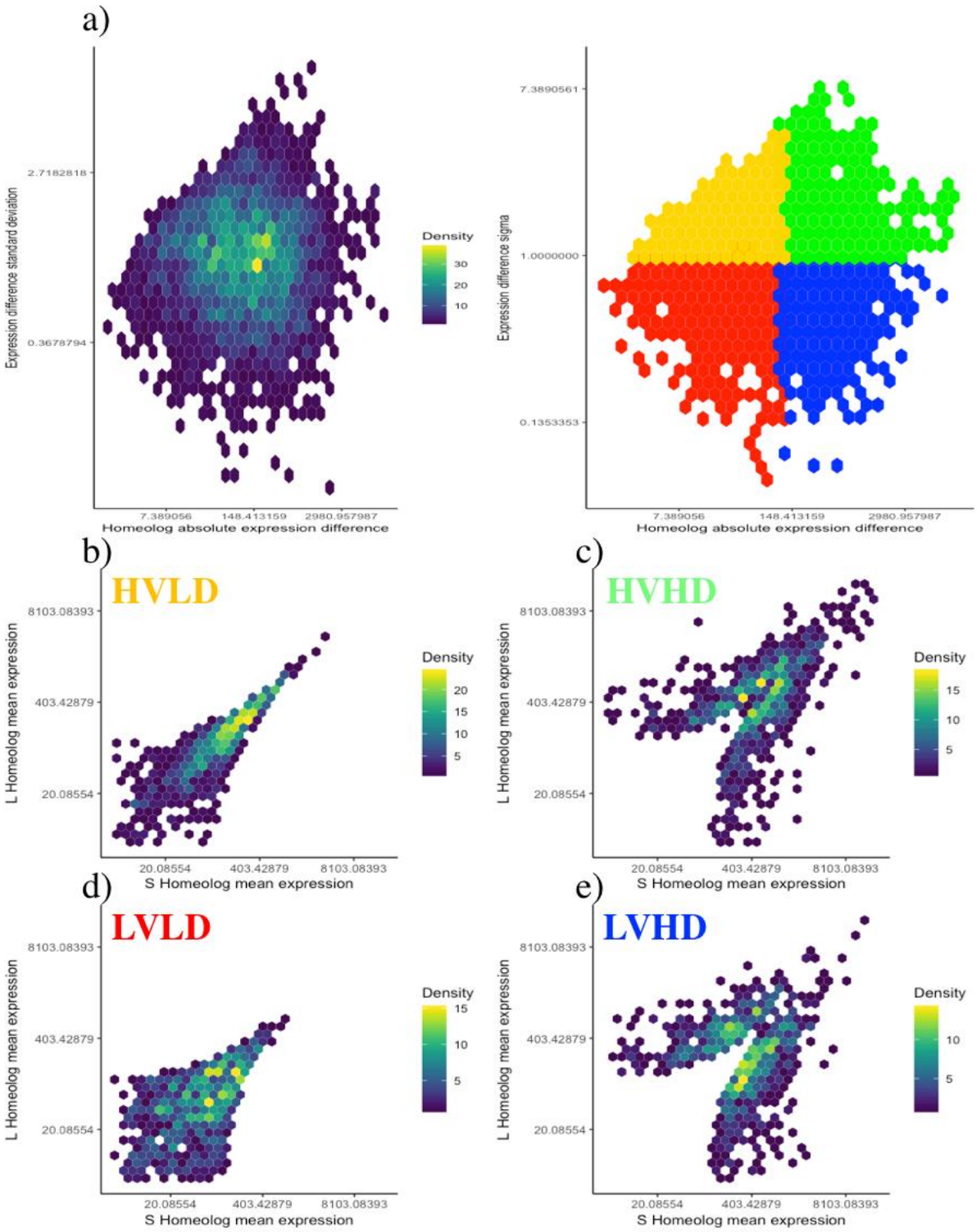


Figure 6. Variance of expression difference and absolute expression difference characterize homeologs into 4 subpopulations at the mid-gastrula stage. a) Initial plot of homeolog absolute expression difference and the variance of the difference was divided into 4 overlapping quartiles with colors corresponding to respective subpopulation. b) High variance of absolute difference and low absolute difference (HVLD) c) High variance of absolute difference and high absolute difference (HVHD) d) Low variance of absolute difference and low absolute difference (LVLD) e) Low variance of absolute difference and high absolute difference (LVHD).

The patterns of homeolog expression within these groups remains relatively constant across all three stages examined and thus we have chosen to use the mid-gastrula stage as our representative example in Figure 6. In the HVLD and LVLD groups, we observe similar expression levels between L and S homeologs and overall lower expression levels of homeologs in both groups. We note a lower amount of correlation between homeolog expression levels in the LVLD group which is unexpected given that the average standard deviation of the expression difference of this group is 0.6, as compared to an average of 1.5 in the HVLD group. This suggest that the relationship between homeologs with diverging expression levels is more stable than those with higher correlation of expression.

The HVHD and LVHD groups are characterized by large differences in expression between the L and S homeologs that diminish as expression increases due to overall expression levels becoming larger than the overexpression differences. The relationship between the divergence of homeolog expression and lower variance in the relationship of homeolog expression is also seen in the LVHD group characterized by greater separation of L and S biased homeologs as compared to the HVHD group (Figure 6e).

We then asked if the grouping of homeologs based on the variation in their expression relationship and differences in their expression levels also grouped them on a functional level. A GO enrichment was performed using the gene sets from each of the four groups against a background of expressed genes at the respective stage. GO enrichments were found to be shared at a low percentage among each of the homeolog groups (Figure 7a), which indicates the relatively specific functional characteristics of each homeolog group which we chose to be represented by the mid-neurula stage. Since many GO enrichment terms are redundant and the list of GO terms produced by each gene set exceeded lengths over 200, we used Revigo (Supek,

Bošnjak, Škunca, & Šmuc, 2011), which employs a clustering algorithm based on semantic similarity to visualize GO terms in a two-dimensional scatterplot. We show the enrichments in the biological process GO category of the homeolog groups at mid-neurula stage in Figure 7.

Functional enrichment of the HVLD group (Figure 7b) resulted in GO terms that represented an overall theme of stress response indicated by the terms: wound healing, DNA damage checkpoint, and regulation of DNA-templated transcription in response to stress. In addition, a tight cluster of GO terms (left) relating to anatomical structure characterized the HVLD group. GO terms in the LVLD group (Figure 7d) appeared to be related to the BMP pathway, autophagy, and histone modifications. However, we noticed a clustering of GO terms which characterized many aspects of neural development such as neural crest cell migration, and autotonomic nervous system development. The HVHD group (Figure 7c) was characterized by ubiquitination, cell division/replication, mRNA processing, and notably developmental patterning as indicated by the anterior/posterior axis specification and axial mesoderm development GO terms that were clustered together. Finally, the LVHD group (Figure 7e) interestingly had GO enrichments relating to mitochondrial functioning such as mitochondrial calcium uptake and oxidation reduction process (not shown) terms. Similar to the HVHD group, mRNA splicing, via splicesasome related homeologs are enriched as well as cellular response to stress which appears to be characteristic to the HVLD group. Notably, we observe a cluster related to epithelial and mesodermal cell fate determination which is unique to the HVLD homeolog group.

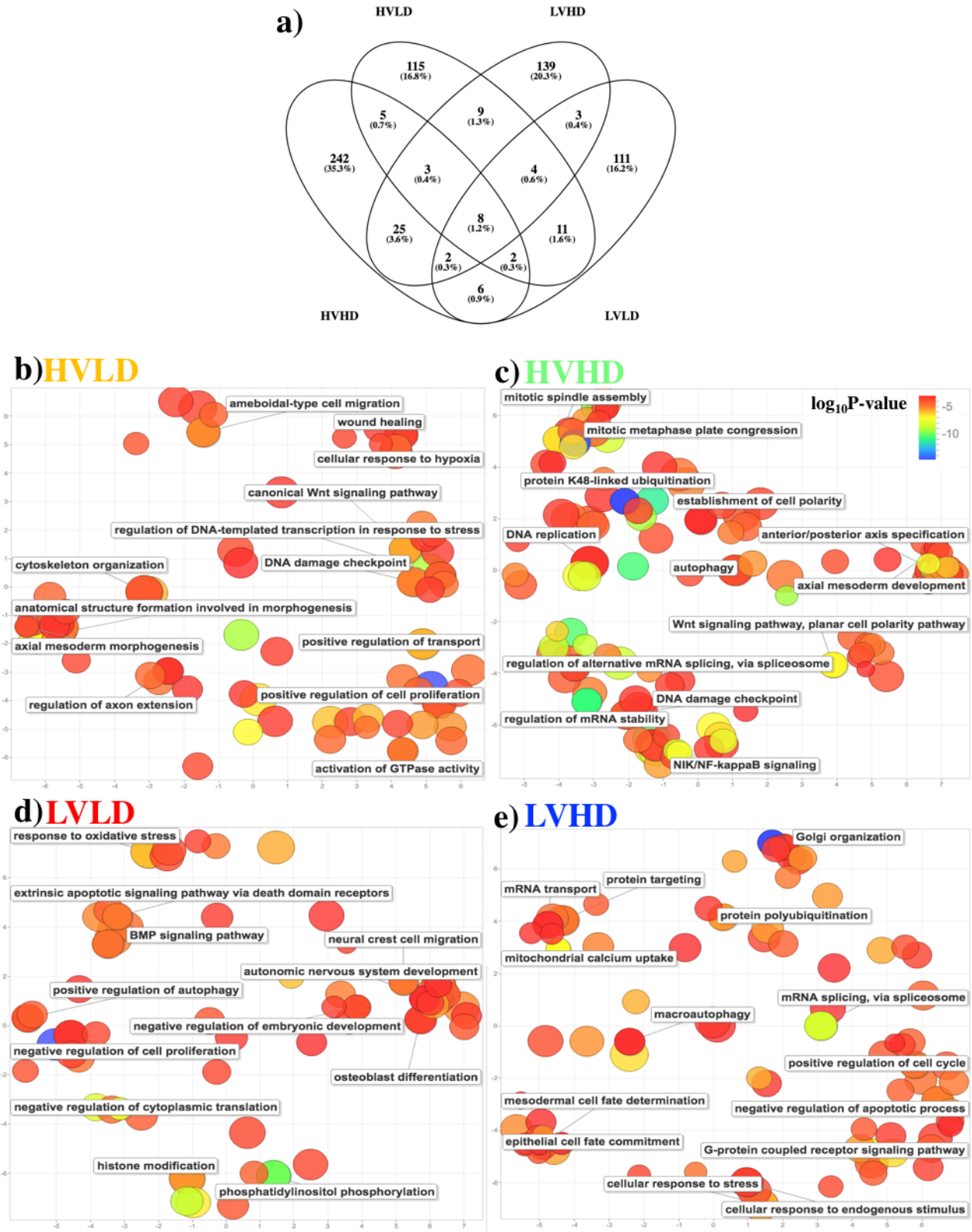


Figure 7. GO enrichment of biological process terms on resulting sets of homeolog genes grouped by variance of expression differences and absolute expression differences. a) A large proportion of GO terms are unique to each homeolog group with the highest amount of sharing occuring between the LVHD and HVHD groups. b) High variance of absolute difference and low absolute difference (HVLD) c) High variance of absolute difference and high absolute difference (HVHD) d) Low variance of absolute difference and low absolute difference (LVLD) e) Low variance of absolute difference and high absolute difference (LVHD). The mid-enurula stage is shown as a representative example. Axis represent semantic space and have no intrinsic meaning. After enrichment, significant ($p < 0.05$) reduced GO terms were filtered and those remaining terms (shown as nodes) were clustered based on semantic similarity using Revigo and relevant descriptions were selected for clusters of semantically similar nodes. Node color represents $\log_{10}$ p-value and size indicates frequency of term in GOA database.

Taken together, these results suggest that homeologs which have a highly variant (HVLD & HVHD) expression differences seem to be specifically involved in DNA damage, morphogenesis, cell division/DNA replication and neural patterning during development. On the other hand, homeologs which exhibit tighter regulation (LVLD & LVHD) appear to be specifically involved in histone modification, cell fate determination, and mitochondria function. Therefore, it appears that both the variance and difference in expression between homeologs plays a role in their functional determination.

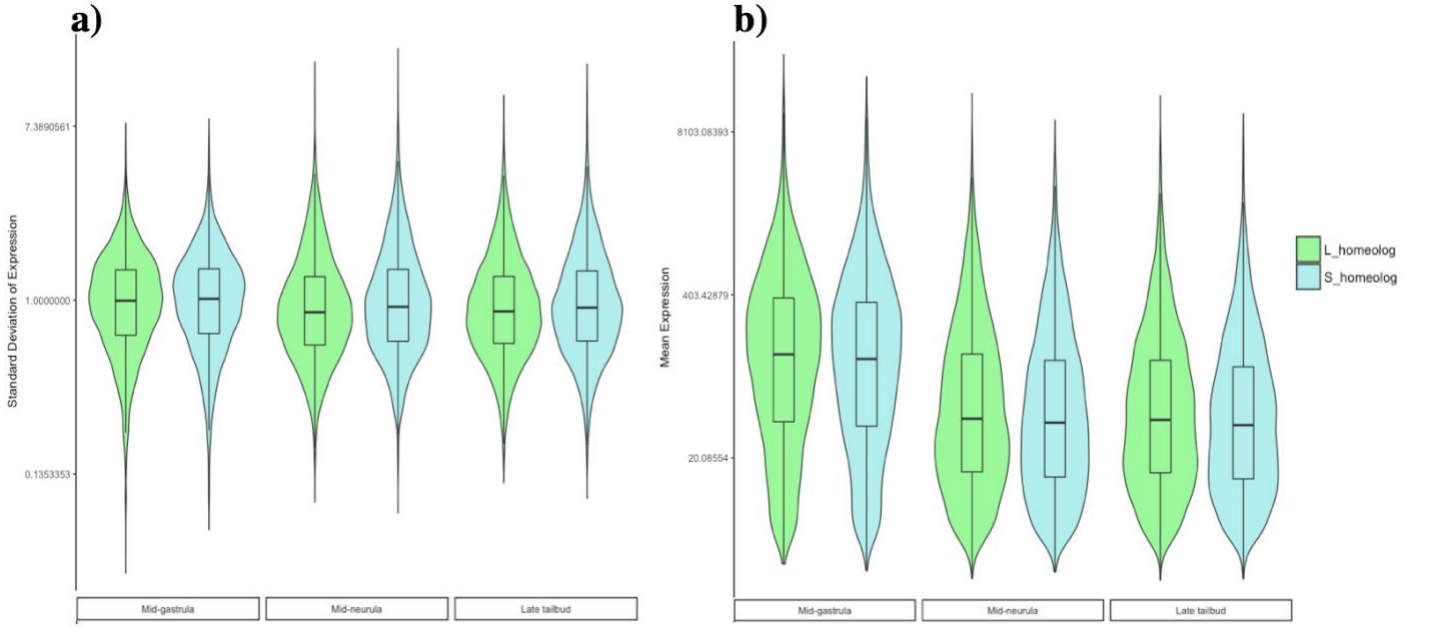
Characterization of Homeolog Expression-Variation Relationship

With results indicating global differences in the variation in homeolog expression, we sought to further characterize the relationship of expression variance with gene expression in homeologs and the expression bias that has been observed between homeologs (Session et al., 2016; Kondo et al., 2017; Michiue et al., 2017; Ochi, Suzuki, et al., 2017; Watanabe et al., 2016) and hinted to in our direct comparison of L and S homeolog expression variability.

We first attempted to use the coefficient of variation as our measure of variation which accounts for genes expressed at different means. However, we observed, as many others have (Bar & Schifano, 2018; Love et al., 2014; Ran & Daye, 2017; Ritchie et al., 2015; Wu et al., 2013; Zhou et al., 2014) that our measure of variance depended on the mean expression and thus examination of expression variance can be confounded by differences in mean expression. To account for the mean-variance relationship, we applied the voom transformation to model the mean-variance relationship which then is used to calculate the standard deviation of expression variance where the mean is accounted for (see methods) and then plot this relationship to confirm to what degree the trend is removed (Figure 9, left). Thus, we used normalized

expression values and a measure of variance that accounts for the mean-variance relationship in this analysis.

On inspection of the relationship of the expression variance between homeologs where both sisters were expressed, we first compared distributions of the mean expression and standard deviation of expression between L and S homeologs. We found a small increase in the total distribution of expression variance that was biased towards the S homeologs at each stage and significantly different from the L homeolog expression variance at mid-neurula and late tailbud



	L Homeolog SD Mean	S Homeolog SD Mean	P-value
Mid-gastrula	0.94	0.96	0.065
Mid-nerula	0.90	0.96	6.55e-10
Late tailbud	0.90	0.94	1.23e-16

	L Homeolog Expression Mean	S Homeolog Expression Mean	P-value
Mid-gastrula	134.8	123.9	0.007
Mid-nerula	41.2	38.3	0.00023
Late tailbud	40.4	36.7	5.33e-06

Figure 8. Global expression variation and mean expression distributions are different between L and S homeologs. a) Expression variation between homeologs is signficantly different at mid-neurula and late tailbud stages, but not at the mid-gastrula stage. Higher expression variation is biased towards S homeologs. b) Mean expression is consistently different between homeologs where there is a bias of higher expression towards L homeologs. Note that the increased mean expression levels at the mid-gastrula stage are due to increased sequencing depth that was unable to be accounted for using the DESeq2 median of ratios method. Comparisons between homeologs are therefore limited to within each stage.

stages (Figure 8a). When comparing this to mean expression values we find that the L homeolog expression means are consistently distributed at a higher levels than S homeolog expression

means which appears to become more asymmetric in later stage, agreeing with previous observation of L homeolog expression dominance (Session et al., 2016). The overall expression means at the mid-gastrula stage were much larger due to higher sequencing depth that was unable be accounted for during median of ratios normalization between all samples. While this prevented direct comparisons across stages, we note that we are still able to make comparisons between homeologs within each stage.

We then asked if there was an association between the expression variation between homeologs despite the overall global differences in the distribution of expression variance. Correlations between the variance of expression at all stages revealed that homeolog expression variance is not strongly correlated between homeologs, while mean expression between homeologs is moderately correlated as expected (Figure 9, right).

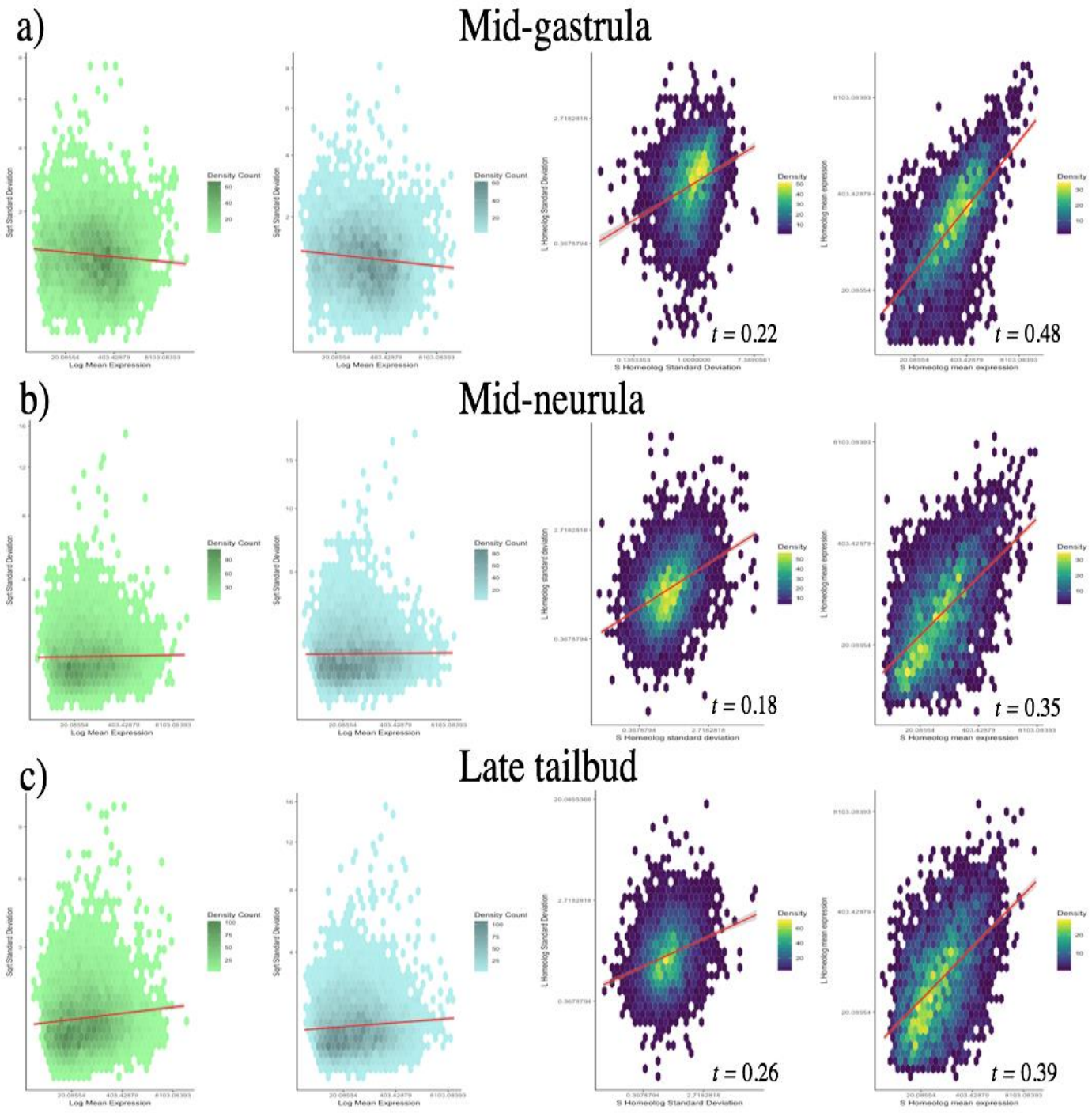


Figure 9. De-trended homeolog variance is weakly correlated between homeologs. The mean-expression relationship shows a weak relationship from fitting a 1st order polynomial line. The minimization of this relationship allows when for comparisons of variation without influence of the mean across all stages (a-c) (left). Comparison of variance between homeologs across all stages (a-c) (right) reveals a weak correlation as compared to the relationship of mean expression between homeologs. Tau is the correlation coefficient using Kendall's rank correlation ($P < 2.2e\text{-}16$)

Differences in Lengths of Homeolog Gene Elements Does not Influence Homeolog Expression Bias or Expression Variance

Large scale genome analysis of the minimal human housekeeping genome and gene expression has previously found correlations between gene expression level and structural parameters such as gene length, exon number, and 3' untranslated region (UTR) (Chiaromonte, Miller, & Bouhassira, 2003). Homeologs differing in gene structure, such as varying lengths in the gene length, 5' UTR length, 3' UTR length, first intron length and exon number should therefore have expression profiles that are influenced on some level by these factors. We tested this hypothesis by first assessing the differences in gene length, 5' UTR length, 3' UTR length, first intron length and exon number between homeologs and then compared whether these differences were correlated with differences in homeolog expression.

The median difference in gene length between homeologs is 7.6Kb which is mostly influenced by the difference in the lengths of the first intron (Figure 10a). Across all 3 stages, mean expression level was moderately correlated with the length. Due to this correlation, correlations between absolute expression difference and length are confounded where decreased absolute difference relies on low mean expression levels that are associated with gene length. No significant correlations were detected between the differences in length between the homeologs compared to differences in homeolog expression or expression variance (Figure 10d, e). An interesting relationship between absolute expression differences and the difference in exon count between homeologs appeared to suggest that larger expression differences stabilized differences in exon counts.

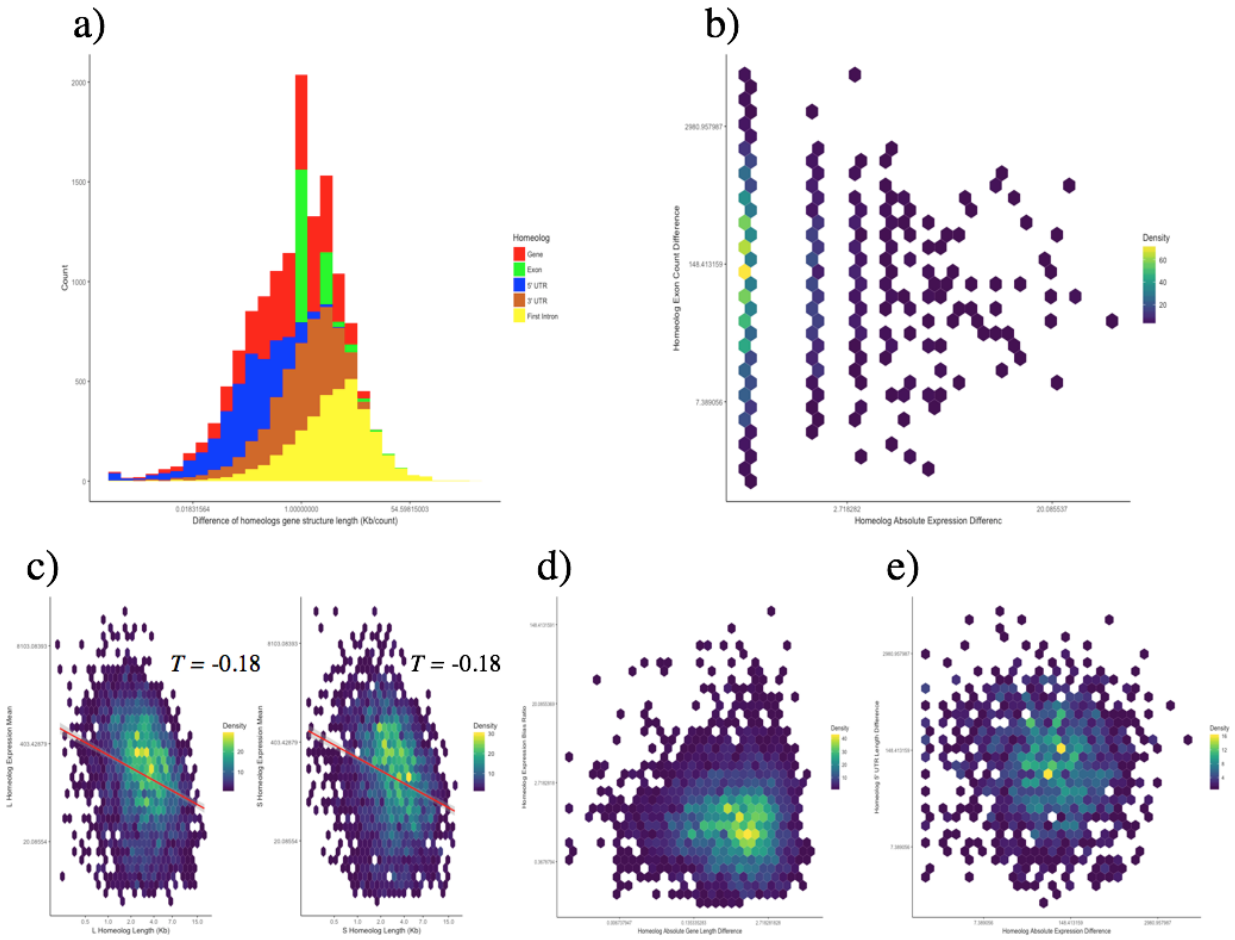


Figure 10. Homeolog expression and expression variance differerences are not influenced by differences in gene structure. Examples are taken from the mid-gastrula stage that is reperesentative of the mid-neurula and late tailbud stages when comparing expression to gene structure. a) Distribution of the differences in length between genes, exon count, 5' UTR, 3' UTR, and the first intron. First intron differences influenced differences in gene length more than other morphlogical parameters examined. b) Increase in absolute expression difference bring the number of exon differences between homeologs to a noisy equilibrium state. Correlation was tested using Kendall's rank correlation test. c) L and S homeolog expression is negatively correlated with increase in gene length. This also implies that as gene length increases, the absolute difference in expression level will decrease. d) Homeolog expression bias expressed as a ratio shows no relationship with the abosulte difference in homeolog gene length. e) The absolute difference in homeolog expression is not correlated wit hthe aboslute difference in 5' UTR regions.

Homeolog Expression Variation is Biased Towards the S Homeolog

As the overall distribution of homeolog expression variance was significantly increased in the S homeologs as compared to the L homeologs (Figure 8a), this suggested that a pairwise comparison of homeolog variance would be able to reveal specific homeolog differential

expression variance. Thus, differential variance analysis was first performed between L and S homeologs within each of the 3 stages using MDSeq (Ran & Daye, 2017). Overall, a larger proportion of S homeologs had significantly greater gene expression variance in the mid-neurula (21% S homeolog bias) and late tailbud (18% S homeolog bias) stages as compared to their respective L homeolog, revealing a bias in variance between homeologs that may only be occurring in later in development. At the mid-gastrula stage, there was no clear variance bias between homeologs where the amount of L homeologs with increased variance was only 6% greater than the amount of S homeologs (Figure 11a).

Differential expression analysis was also performed between homeologs to see if there was any association between differentially expressed genes and differentially variant genes. On average, 85% of homeologs that exhibited differential variance were also differentially expressed (Figure 11c). This overlap was surprising as the software that was used to test for differential variance is supposed to account for the mean-variance relationship. To confirm that this overlap was not confounded by the mean-variance relationship, we identified homeologs contained in this overlap where the increased variance occurred concurrently with increased expression. Homeologs that met this criteria exhibit the opposite trend that is expected of the mean-variance relationship and thus not confounded by this. Out of these the homeologs biased towards increased expression variance, 91% of L homeologs and 97% of S homeologs were also positively differentially expressed, respectively (Figure 11d). Thus, we conclude that there the increased variance bias observed in homeologs is not confounded by expression means.

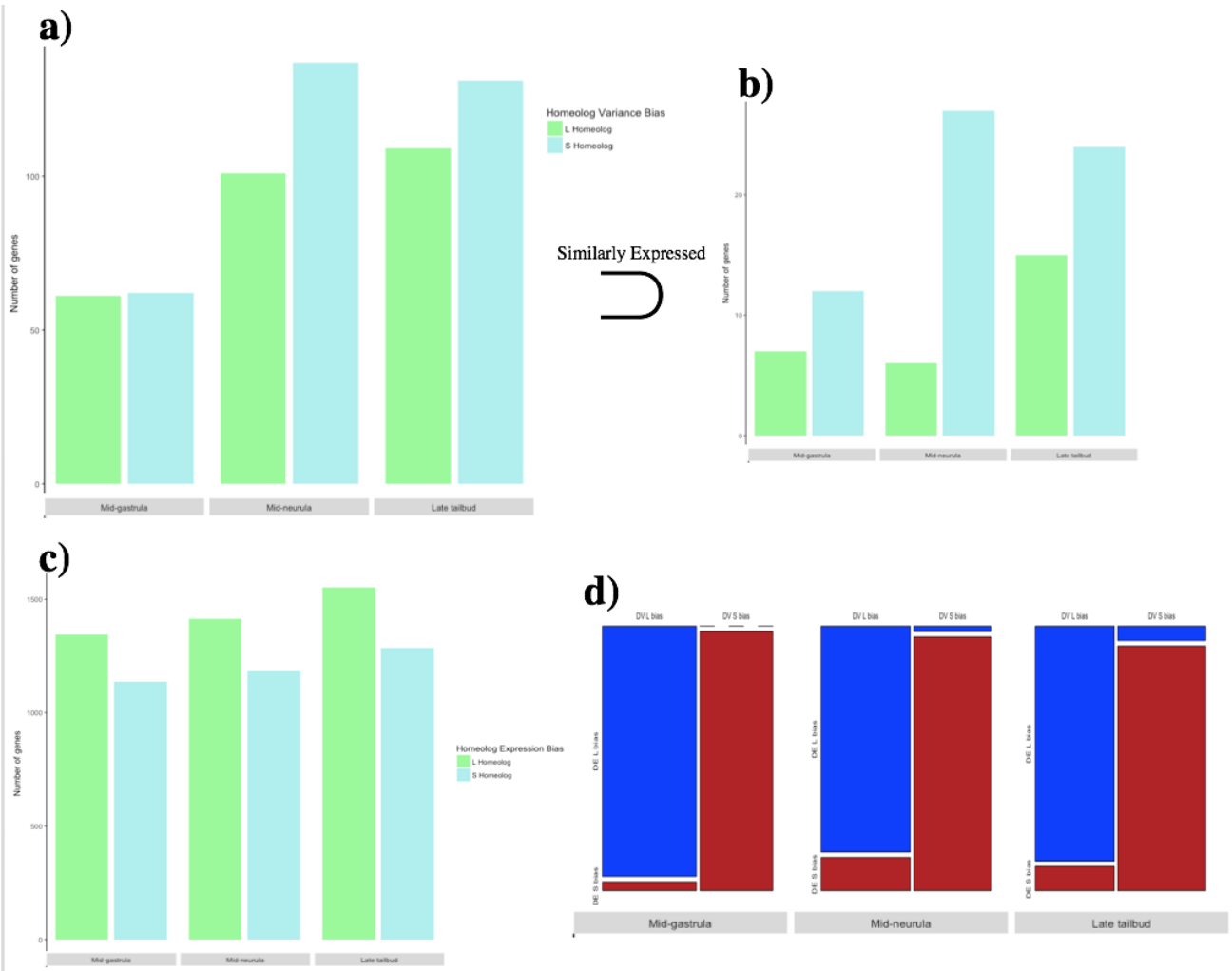


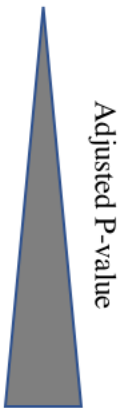
Figure 11. S homeologs exhibit expression variance bias. Expression variance between homeologs was tested at 3 developmental stages where homeolog expression variance occurs at all stages differentially biased at mid-neurula and late tailbud stages. a) Homeolog variance bias is defined as an increase in expression variance. Expression variance is not biased at the mid-gastrula stage, but becomes biased towards S homeologs at the mid-neurula and late tailbud stages later in development. b) The subset of genes that were similarly expressed but where there was differential expression variance. Across all stages bias is towards the S homeolog and is as much as 3-fold higher than the L homeologs at the mid-neurula stage. c) Differential expression between homeologs where the differential expression is biased towards the L homeolog meaning that the expression levels are greater than the S homeolog. d) Mosaic plots of the overlap between genes that are both biased in differential expression (DE) and differential expression variance (DV) towards either the L or S homeolog. If there is 100% agreement in shared trends, than the left columns will be solidly blue and right columns will be solidly red. There is high agreement in homeolog specific bias between the shared trends of DE and DV within each homeolog. This indicates that a relative increase in variance is associated with a relative increase in expression which is not confounded by the mean variance relationship.

While the majority of homeologs with differential expression variance where also differentially expressed, a small subset of these were similarly expressed as determined by an insignificant result from differential expression testing. Out of these similarly expressed

homeologs, the number homeologs with increased expression variance over 50% greater than the L homeologs (Figure 11b). The homeolog specific increases in variance suggest that S homeologs may either be dysregulated in response to asymmetric pseudogenization biased towards S homeologs, or are important in the robustness and plasticity during embryonic development.

Gene ontology (GO) enrichment of differential variance homeologs was performed against a background of expressed genes at each stage to gain more insight on the functional processes these gene sets are related to. This resulted in a large proportion of many genes related to protein transport, localization and modification across all 3 stages. Another common theme among GO terms was cellular organization which also was enriched across all three stages. In the cellular component category of GO terms, DV homeologs at the mid-gastrula stage were highly enriched in intracellular components (39% of all DV) while DV homeologs at mid-neurula and late tailbud stages were highly enriched in extracellular exosome components (17% & 16% of all DV, respectively. Separate enrichments of variance biased S homeologs revealed were specific to protein

Mid-gastrula	Mid-neurula	Late tailbud
Organelle organization	Protein transport	Anatomical structure morphogenesis
rRNA methylation	Regulation of response to stress	Protein transport
Cellular protein modification process	Cell component organization	Regulation of cellular component organization
Apoptotic signaling pathway	Regulation of protein localization	Protein localization to organelle



21-25% 16-20% 11-15% 6-10% 1-5%

Table 1. GO enrichment of differential expression variance homeologs. Enrichment was performed by stage using against a background of expressed genes respective to the stage and reduced to most specific terms. Only Top 4 biological process GO terms are shown. Shading indicates percentage of homeologs in a GO category out of total differntial variance homeologs within each stage. Most highly significant GO terms are towards the top of the table.

transport at the mid-neurula and late tailbud stages. At the mid-gastrula stage, variance biased S homeologs enrichments were specific apoptotic signaling pathway and organelle organization.

Homeolog Expression Bias Response to Genetic and Physical Perturbations

The overall theme of the Saha Lab is to study the robustness of embryos in response to chemical, physical, and genetic perturbations during embryonic development. The major interest is to elucidate the mechanisms of the compensatory responses to these perturbations that allow the developing embryo to recover and continue to proceed with development. We have approached this question from a transcriptomics perspective which has allowed us to observe the transcriptional changes in response to genetic or physical perturbations.

Previously, we have identified differentially expressed genes following the response to perturbations over several time points. Among these differentially expressed genes, on average 67% are homeologs. Functional analysis of the differences in homeolog gene retention has suggested that many developmentally important signaling pathways such as Notch, TGFB, Wnt, Hox, Hippo and Hedgehog (Session et al., 2016) are retained at higher rates than the expected rate for other homeologs. This higher retention rate is also associated with a small bias ratio during development (oocyte-stage20) that later becomes biased towards predominately biased towards L. This suggest that since homeologs already play an important role in developmental processes and that L homeolog subgenome dominance is not realized until later stages, homeologs may be involved in the response to perturbations during development as well as the compensatory mechanisms that restore normal developmental phenotypes. Therefore, differential homeolog bias in response to perturbation may indicate homeolog subfunctionalization in compensatory response where the non-dominant homeolog plays an important role in this

response. This may also be the case where perturbation induces homeolog bias where normally there is none. On the other hand, differential homeolog bias may be due to cis-regulatory structural differences where one of the homeologs has become pseudogenized and lost the ability to respond signaling factors that are part of the initial response to a perturbation or the compensatory response.

Homeolog bias in response Notch Genetic Perturbations

To study genetic perturbations during embryonic development, the Notch signaling pathway was either hyperactivated or inhibited using two genetic constructs. Briefly, a genetic construct notch which coded for the *Xenopus* laevis Notch Intracellular Domain (ICD) was used to hyperactivate the notch pathway while a DNA binding mutant (DBM) was used to inhibit Notch signaling by sequestering transcription factors associated with the CSL (CBF1, Suppressor of Hairless, Lag-1) transcription factor complex (Pursglove & Mackay, 2005). This relatively simple means of perturbation was performed unilaterally at the 2-cell stage, where the control embryos were unilaterally injected with a GFP construct. Following the initial perturbation, mid-neurula, early tailbud, and late tailbud (Nieuwkoop & Faber, 1994) were stages selected to study the compensation from this perturbation over a time-course. RNA-Sequencing was performed at each of these stages following the initial perturbation using a pooled sample of 5 embryos.

Preliminary analysis of differentially expressed genes between experimental (ICD & DBM) compared to control (GFP) conditions at three different stages after the initial perturbation revealed that on average 66% of differentially expressed genes are homeologs. This is higher proportion of homeologs than is expected as homeologs make up 23.3% of the total gene in *X. laevis*. Since we and others (Session et al., 2016; Chain & Evans, 2006; Kondo et al., 2017;

Michiue et al., 2017; Watanabe et al., 2016) have observed expression bias towards the L homeolog, we first checked whether this expression bias was present in our samples. We used DESeq2 to directly compare expression levels of L and S homeologs where we found L homeologs had significant expression bias where the amount of L homeolog expression bias was 7.3% more than S homeologs expression bias (Figure 12a). Getting a fraction of the L homeolog expression over the S homeolog expression per homeolog, we found differing amounts of homeolog expression bias among the conditions at all three stages. Interestingly, this revealed a bimodal distribution among all samples indicating that homeologs tend to be biased and not evenly expressed (Figure 12b).

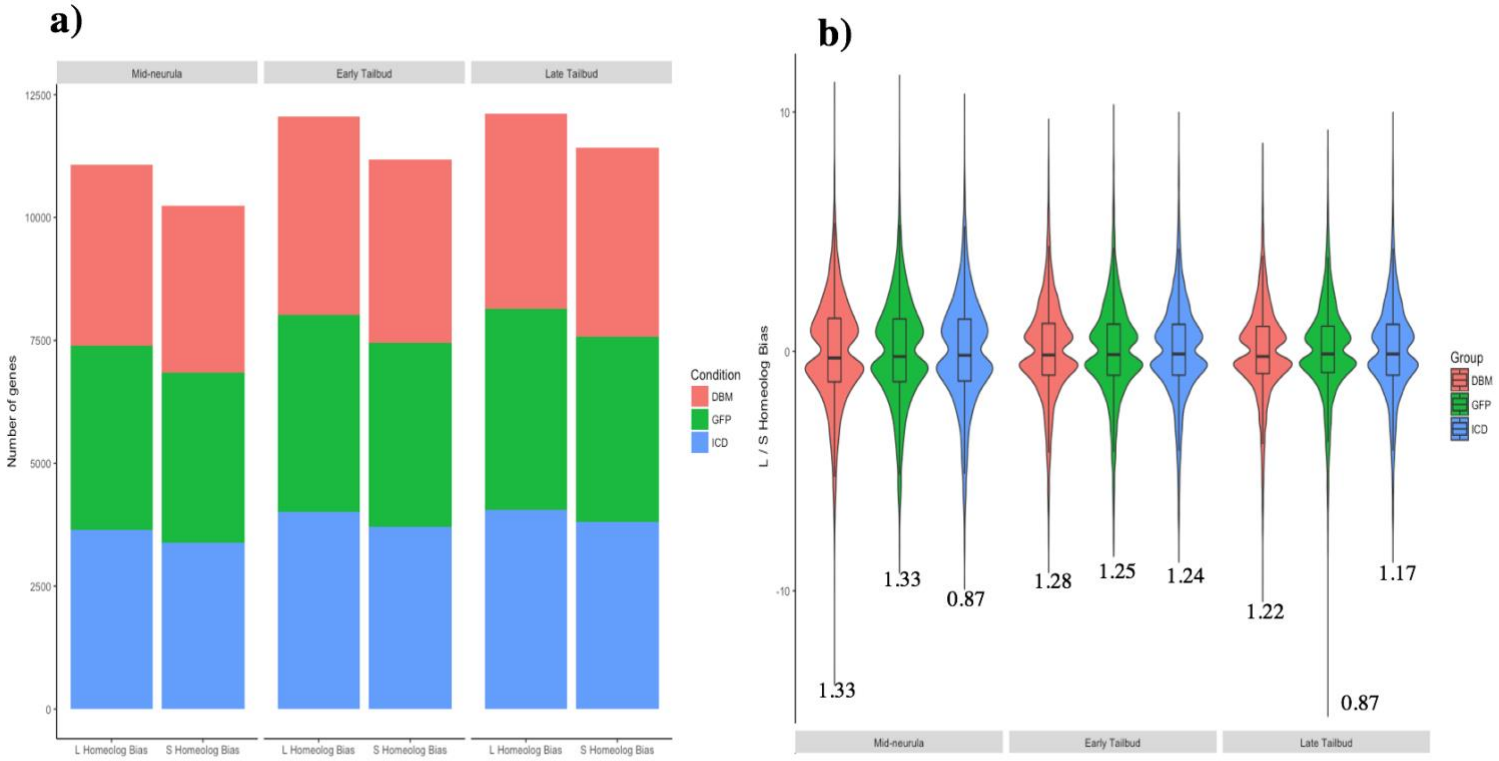


Figure 12. Homeolog expression in genetically perturbed embryos is biased towards L homeologs. a) Number of homeologs exhibiting expression bias greater towards the L homeolog in all conditions across all stages, which agrees with previous observations (Session *et al*. 2016). b) Ratio of homeolog bias shows a bimodal distribution of either L or S bias where the density is lowest near zero, indicating the majority of expression ratios are biased.

From our observation of globally varying homeolog expression levels between samples, we next asked if a direct comparison between the homeolog biases across conditions would reveal any changes in the relationship between homeologs that are induced by perturbations and potentially involved in the compensatory response. We tested this with DESeq2 where we modified the design matrix for the general linear model to estimate the ratio2 of homeolog expression in each condition while accounting for variability within samples and differing gene lengths between homeologs (see methods). On comparison of the control condition (GFP) with the experimental conditions (ICD, DBM) at each stage respectively, we found 4 different patterns of differential homeolog bias.

We observed the most common change in homeolog bias across all comparisons to be an increase in bias, meaning that the difference between expression levels between homeologs had significantly increased (Figure 13a). Following this was a decrease in bias, meaning that expression levels between homeologs had significantly decreased. Within both patterns of changes in bias, the proportions of L and S homeologs were relatively similar, indicating no homeolog specific preference. Switches in homeolog bias either from L bias to S bias or vice versa were a rare occurrence that was most prevalent at the early tailbud stage.

Since differentially biased homeologs implied changes in expression between the two conditions being compared, we looked at the overlap of the differentially biased homeolog gene set against significantly the sets of biased homeologs and all differentially expressed genes. We found a the largest proportion of differentially biased homeologs overlapped with the biased homeolog gene set at all comparisons (Figure 13b). Despite relatively low amounts of differentially homeolog biased genes in the DBM comparisons as compared to the ICD

comparisons at the early and late tailbud stages, a larger proportion of differentially biased homeologs with unique to the differential bias comparisons was prevalent which may indicate that changes in homeolog bias are transcriptionally important in the recovery from inhibition of notch signaling.

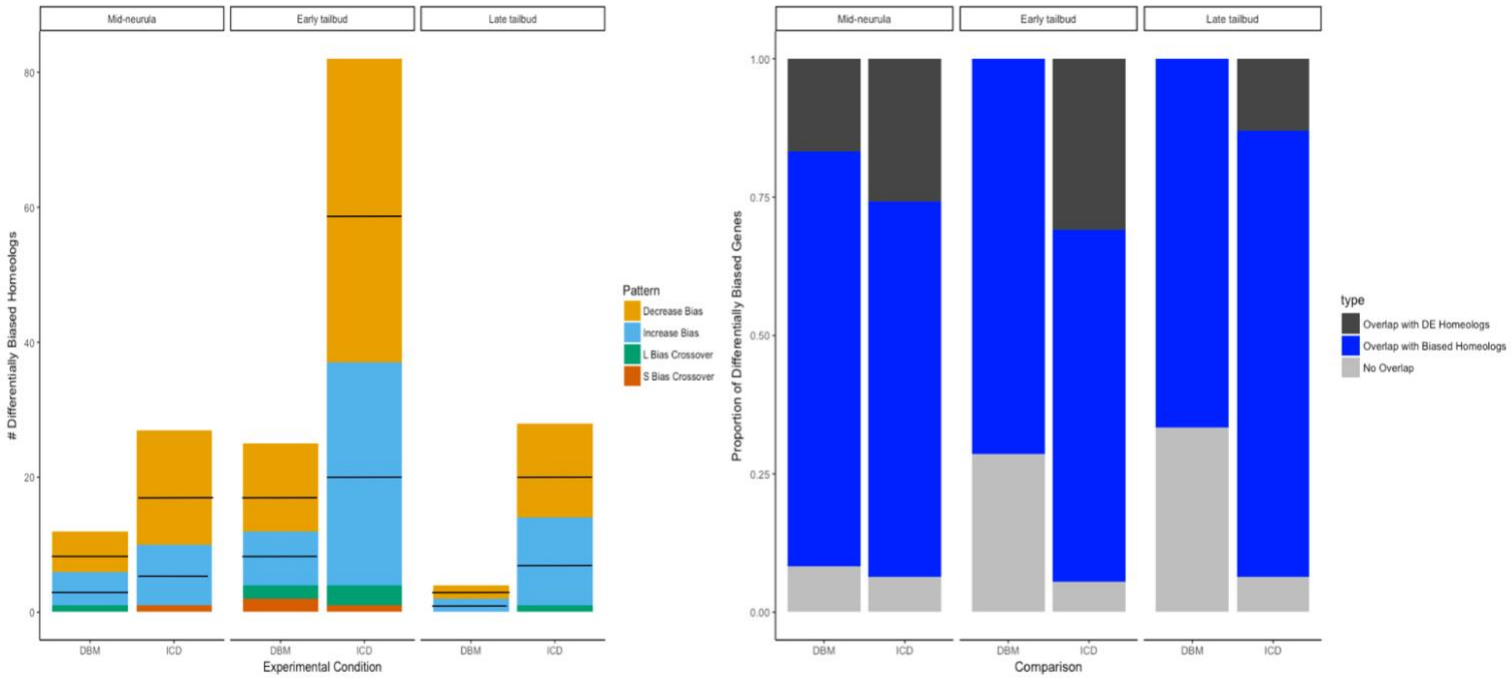


Figure 13. Perturbation induces differential homeolog bias. a) Total count of differentially biased homeologs resulting from the comparison of the GFP condition (control) to the ICD or DBM (experimental) conditions respective to each stage. Differential bias is further categorized into 4 different patterns (bar plot fill) that describe that specific bias change. Within the 'Decrease bias' and 'Increase bias' patterns, a line is drawn to indicate whether the increase in bias was towards the L (top) or S (bottom) homeologs, respectively. b) Proportion of overlap between the differentially biased homeolog gene sets and the differenetially expressed genes (dark grey) or biased homeologs (blue). No overlap indicates homeologs that were unique to the differentially biased gene set (light grey). Overlap between differentially biased homeologs was greatest between biased homeologs in all comparisons indicating the majority of homeologs that are differntialy biased are biased within each experimental condition.

Homeolog Expression Bias in Response Anterior-Posterior Neural Axis Rotations

To study physical perturbations during development we disrupted patterning of the anterior-posterior (AP) axis during development of nervous system in *X. laevis*. Experiments performing a transplanting the AP neural axis tissue from a donor embryos, rotating this 180-degrees and then incorporating this into a host embryo that has had the same section of AP neural axis tissue removed. This experiment was performed during gastrulation and extensively

perturbed the patterning in the early development of the central nervous system. Interestingly, embryos are able to compensate from this perturbation by the hatching stage, which represents the ability to re-pattern the AP axis following an inversion.

In order to investigate the role of the differential gene expression in these compensatory responses, RNA-Seq was performed on whole embryos at mid-neurula and late tailbud stages (Nieuwkoop & Faber, 1994) following the initial perturbation using 5 biological replicates in each condition. Experimental conditions were defined either as embryos that underwent transplantation surgery during gastrulation (rotated), embryos that underwent transplantation surgery but the transplant was not rotated (sham) and embryos that underwent transplantation surgery where the donor and host embryo were the same embryo (autotopic). The autotopic condition was included to attempt to control for the confounding effect of a xenobiotic piece of tissue from the donor embryo being transplanted to the host. In addition, embryos in the same batch as experimental embryos that did not experience any exposure to experimental conditions or control conditions (sibling) were included to test how different the sham embryos were due to transplantation alone. In total, this gave us RNA-Seq profiles of 4 different conditions to go beyond differential gene expression analysis and look at the homeolog bias patterns as was identified in the Notch perturbations.

From the observations of homeolog bias and changes in homeolog bias observed in the genetic perturbation to the notch signaling pathway, we hypothesized that physical perturbation during embryonic development might induce similar patterns of homeolog expression that might connect the observations between these two perturbation experiments.

Throughout the analysis of the RNA-Seq data we focused on 1) comparisons between the sibling group to either the sham or autotopic; 2) comparison between the sham group to either

the rotated or autotopic groups. Following standard differential gene expression analysis, of the aforementioned comparisons, we found on average 69% of differentially expressed genes to be composed of homeologs which is similar to the proportion found in the Notch experiments. Next, we checked the expression bias between homeologs using DESeq2 which yielded a greater proportion of L homeolog bias genes as expected and also observed in the genetic perturbation samples (Figure 14a). Plotting the homeolog bias ratios for each sample revealed a bimodal distribution where the density of S biased homeologs (negative ratios) appeared to have the greatest density near 0, indicating that the magnitude of S homeolog expression bias is relatively low as compared to the wider distribution of L homeolog bias ratios.

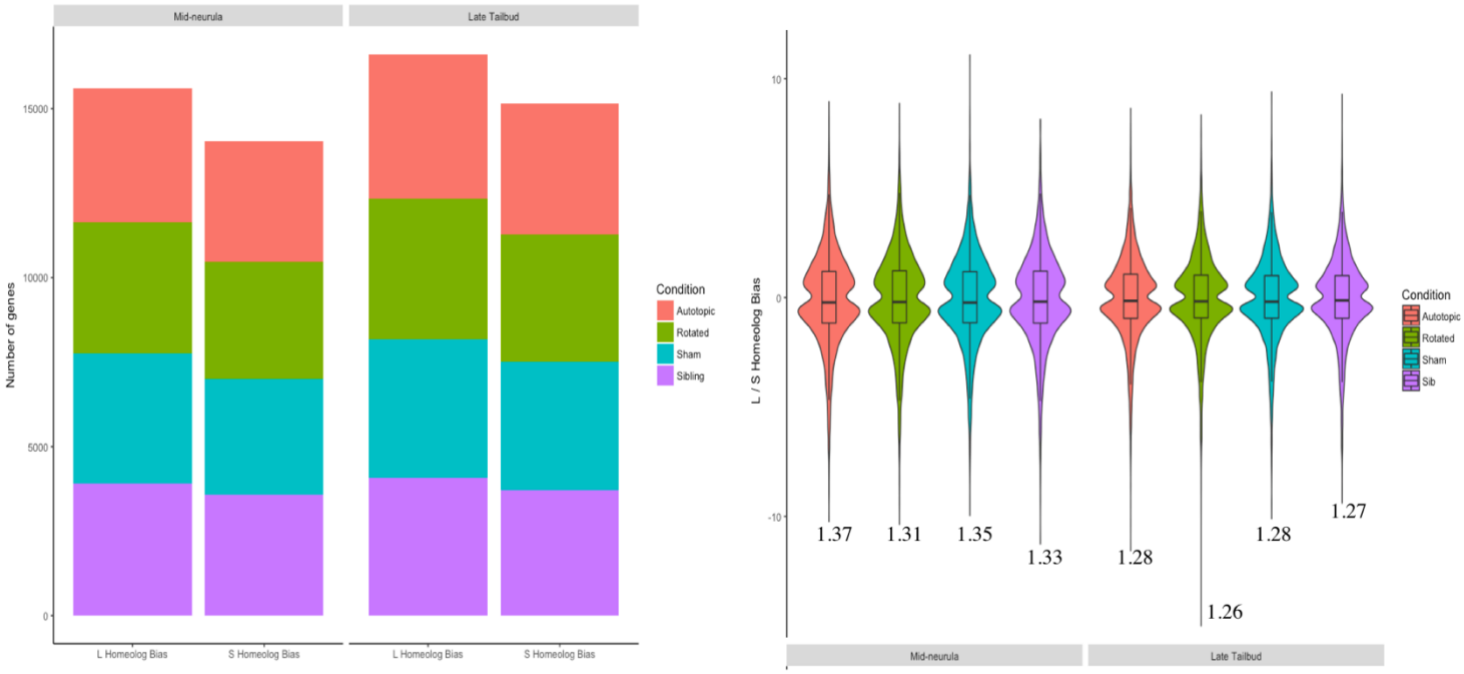


Figure 14. Homeolog expression in physically perturbed embryos is biased towards L homeologs. a) Number of homeologs exhibiting expression bias greater towards the L homeolog in all conditions across all stages, which agrees with observations from genetically perturbed embryos. b) Ratio of homeolog bias shows a bimodal distribution of either L or S bias where the density is lowest near zero, indicating the majority of expression ratios are biased.

We then checked the patterns of changes in homeolog bias among the comparisons to see if there was any differential bias between homeologs. We observe an overall decrease in the

amounts of differentially expressed homeolog from the mid-neurula to the late tailbud stages indicating that the response to the physical perturbation has diminished later in development (Figure 15a). Since these comparisons focused on embryos which compensated from physical perturbations by the hatching stage, this suggest that the overall compensatory mechanisms are also diminishing due to restoration of the normal phenotype. Comparisons of Sham vs Autotopic and Sibling vs Autotopic resulted in the greatest amounts of differentially biased homeologs across both stages while comparisons of Sham vs Rotated and Sibling vs Rotated resulted in 95% less differentially biased homeologs on average (Figure 15a). In contrast to the relatively even amount of increase/decrease in homeolog bias observed in the genetic perturbations, we see a large proportion of increases in bias that account for 48% and 58% of the total changes in homeolog bias at mid-neurula and late tailbud stages, respectively (Figure 15a).

Since our main interest in these physical perturbations was to screen for genes, in this case homeologs which show changes in homeolog bias, related to the re-patterning of the AP neural axis after a rotation, we reasoned that a comparison among the resulting gene sets from each comparison could give us insight on the changes in homeolog bias unique to recovering from the AP neural axis rotation. We performed four-way comparisons among gene sets within each stage. At the mid-neurula stage, many genes were shared between the comparisons of the Sham and Sibling conditions against the autotopic condition while no homeologs were shared among all 4 comparisons (Figure 15b). 8 homeologs remained exclusive to the Sham vs Rotated comparison which either may be important in the compensation process (Figure 15b). Similarly at the late tailbud stage, all 6 differentially biased genes in the Sham vs Rotated comparison are not shared among other comparisons (Figure 15c) and also do not overlap with the 8 homeologs exclusive to the same comparison at the mid-neurula stage. This difference between the same

comparison at a later point in the time course of recovery may indicate time specific changes in homeolog bias that play a role in compensation.

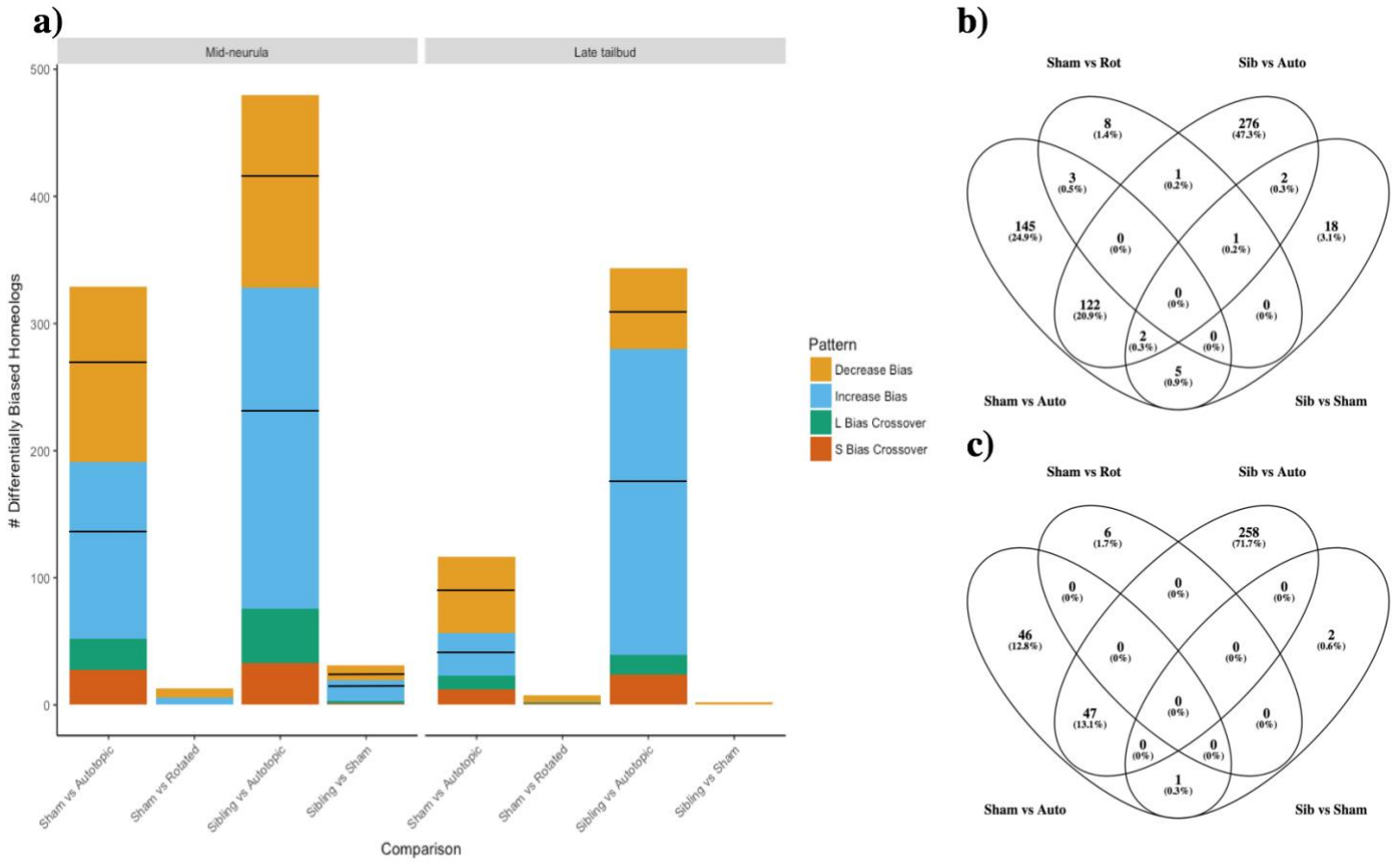


Figure 15. Phsyical perturbation induces differential homeolog bias. a) Total counts of differentially biased homeologs resulting from parallel comparisons made at the mid-neurula and late tailbud stages. Differential bias is further characterized into 4 different (bar plot fill) that describe specific changes in bias. Within the 'Increase Bias' and 'Decrease Bias' patterns, a line is drawn to indicate the proportion of L homeologs (top half) and S homeologs (bottom half). 4-way overlaps were made between all comparisons made at the mid-neurula b) and late tailbud c) stage. Counts of homeologs are indicated at each overlap with an acompanying percentage out of the total amount of homeologs.

Discussion

Limitations

While variance analysis of homeologs uses robust statistical techniques (Ho, Stefani, Dos Remedios, & Charleston, 2008; Ran & Daye, 2017) that are similar to the current techniques used for differential expression analysis (Love et al., 2014; Ritchie et al., 2015; Zhou et al., 2014), it is largely confounded by small sample sizes that do not reveal the true variability amongst individuals. As this study expanded on the findings (Session et al., 2016; Michiue et al., 2017; Watanabe et al., 2016) that only used a sample size of 2 and increased this to sizes of either 5 or 8, larger sample sizes are needed in order to acquire robust results that are proven to be repeatable across multiple studies.

We are constrained to comparing many measures of homeologs to an expectation and do not have any direct comparison available. This is because direct comparison of homeologs and non-homeologs is confounded by the fact that non-homeologs are composed of different genes and gene sets that seem to have different behaviors than homeologs. Here, we have attempted to avoid this problem by constructing null expectations using repeated permutation tests in order to see if there are significant differences in homeologs as compared to this expectation. An alternative to this is to make direct comparisons of homeologs with their orthologs in the diploid *Xenopus tropicalis*, which has previously been done in the literature (Peshkin et al., 2015). While this removes the confounding effect of gene composition difference, it introduces countless other effects such as overall magnitude differences in transcription due to cell size differences as well as the lack of non-homeolog genes that may be interacting with homeolog genes.

The amount of data is quickly surpassing the rate of analysis. As a result, online repositories have become a goldmine for large scale data analysis. However, every experiment

and data from these experiments are associated with batch specific effects which can heavily confound analysis if not properly accounted for. Here, we pooled data from 5 different studies in order to increase the sample size of our analysis. However, we also noticed effects specific to each experiment in terms of sequencing depths and other artifacts that do not allow unbiased analysis between these samples. We attempted to correct for this using the statistical frame work of linear models employed by the limma (Ritchie et al., 2015) and DESeq2 (Love et al., 2014) packages, however we still noticed batch effects between stages (Figure 8b). While there are a handful of tools (Liu & Markatou, 2016; Zhou et al., 2014) which aim to correct for batch effects, a better solution is needed in the age of big-data analysis.

Homeolog variance

This study set out to further characterize highly variable homeolog expression during embryonic development and gain insight on the mechanisms that underlie this. Here we found that individual homeologs exhibit a greater amount of gene expression variance than non-homeologs. Since homeolog pairs are, by definition, very similar in their coding regions, their gene products must also be very similar functionally. Moreover, the function or differentiation between gene products from either homeolog pair are indistinguishable. Thus, while homeologs may exhibit high expression variance individually, their combined expression levels that represent the indistinguishable level of a gene product should be at less variant than the variance of individual homeologs. In support of this, we have provided evidence for this hypothesis where we find that the sum of homeolog expression does indeed have an overall lower variance than individual homeologs. However, further study needs to make comparisons of the expression sum variance beyond the null expectation.

On the other hand the high variance observed between homeologs as well as in the relationship between homeolog expression might be indicative of a loss of cis-regulatory elements that is specific to one of the homeologs. Our results suggest that expression variance is biased towards S homeologs. Since the rate of gene loss between homeologs has been shown to be asymmetrically skewed towards S homeologs (Session et al., 2016), this suggest that homeolog specific gene expression variance might be associated with this loss. It has been shown in *X. laevis* through transgenic experiments (Ochi, Kawaguchi, et al., 2017; Ochi, Suzuki, et al., 2017) that cis-regulatory elements drive specific levels and patterns of expression. Thus, it remains possible that mutations in these regions due to relaxed selective pressure can cause gene expression variance, which lacks sufficient study.

To explore this, we attempted to make preliminary associations between the lengths of gene elements and the expression bias or expression variance bias, but did not find any significant correlations. Such simple correlations are thus not strong enough to explain these differences and must be approached by direct comparison of regulatory sequences. This explanation of expression variance can be approached in-silico by a large scale association of expression variance with the rate of mutation in important cis-regulatory regions and is the interest of future study. In addition, attention in the polyploidy field has recently turned to role of epigenetic mechanisms (Jackson & Chen, 2011) in explaining differences in expression changes and may also explain expression variance. Thus, future studies need to also utilize epigenetic data to overlay this with sequence and expression data in order to better explain both differences in expression but also differences in the variance of expression.

Homeolog expression bias in response to genetic and physical perturbations

The process of development must be robust to account for both intrinsic and external perturbations that threaten the survival an already vulnerable embryo. Complex genetic networks must be able to coordinate compensatory responses to a wide range of different types of perturbations. Genetic redundancy has been observed in signal transduction and transcriptional networks that may provide a buffer to any changes in the signaling genes or perturbations to the network itself (Wagner & Wright, 2007).

The allotetraploidy event in *Xenopus laevis* was a merger between two diploid species that has left many duplicate copies of functioning genes. Around 46% of the genome is still duplicated where a global functional analysis of these has indicated that these are genes are enriched in developmental processes (Session et al., 2016). Thus, in the context of developmental perturbations, homeologs may provide an important role which allows for allotetraploid embryos to better compensate from these.

From RNA-Sequencing results over a time course following initial perturbations, we were able to inquire about the response of homeologs to either genetic or physical perturbations. We find that changes in homeolog bias are associated with both types of perturbations, but the patterns by which the changes in homeolog bias occur seem to be specific to the mode of perturbation. This analysis has provided several candidate genes on which to investigate their role in the response to perturbation. The most interesting of these are the homeologs which switch their bias, which may suggest homeolog novel homeolog specific functions.

While these results confirm that homeologs bias changes in response to perturbation during development, it cannot address the absolute importance of these changes and whether they or necessary or sufficient for the recovery process. While it has been suggested in the

literature that This calls for studies using species of *Xenopus* with varying ploidy from the diploid *X. tropicalis* up to the dodecaploids *X. longpipes* and *X. ruwenzoriensis* which can elucidate the role of increasing ploidy, and thus a larger genetic buffer, on developmental robustness. Now that the *X. laevis* genome has been successfully assembled, this can be used as a reference for other *Xenopus* assemblies in order to perform large scale genetic studies. However, this is currently limited to the genome and transcriptome assembly technologies (Duan, Xia, Zhao, Jia, & Kong, 2012; Krasileva et al., 2013; Nakasugi, Crowhurst, Bally, & Waterhouse, 2014) that are confounded by not only long repetitive sequences, but highly similar homeolog sequences that must be separated in the assembly process.

References

Altmann, C. R., & Brivanlou, a H. (2001). Neural patterning in the vertebrate embryo. *International Review of Cytology*, *203*, 447–482. http://doi.org/10.1002/dvdy.20464

Andersson, E. R., Sandberg, R., & Lendahl, U. (2011). Notch signaling: simplicity in design, versatility in function. *Development*, *138*(17), 3593–3612. http://doi.org/10.1242/dev.063610

Bar, H., & Schifano, E. D. (2018). Differential variation and expression analysis, 1–13.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., … Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, *5*. http://doi.org/10.1038/ncomms4657

Bottani, S., Zabet, N. R., Wendel, J. F., & Veitia, R. A. (2018). The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Trends in Plant Science*, *xx*, 1–10. http://doi.org/10.1016/j.tplants.2018.01.002

Breckenridge, R. A., Zuberi, Z., Gomes, J., Orford, R., Dupays, L., Felkin, L. E., … Mohun, T. J. (2009). Overexpression of the transcription factor Hand1 causes predisposition towards arrhythmia in mice. *Journal of Molecular and Cellular Cardiology*, *47*(1), 133–141. http://doi.org/10.1016/j.yjmcc.2009.04.007

Chain, F. J. J., & Evans, B. J. (2006). Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog Xenopus laevis. *PLoS Genetics*, *2*(4), 478–490. http://doi.org/10.1371/journal.pgen.0020056

Chiaromonte, F., Miller, W., & Bouhassira, E. E. (2003). Gene Length and Proximity to Neighbors Affect Genome-Wide Expression Levels Gene Length and Proximity to Neighbors Affect Genome-Wide Expression Levels. *Genome Research*, 2602–2608. http://doi.org/10.1101/gr.1169203

Cox, M. P., Dong, T., Shen, G. G., Dalvi, Y., Scott, D. B., & Ganley, A. R. D. (2014). An Interspecific Fungal Hybrid Reveals Cross-Kingdom Rules for Allopolyploid Gene Expression Patterns. *PLoS Genetics*, *10*(3). http://doi.org/10.1371/journal.pgen.1004180

Ding, Y., Ploper, D., Sosa, E. A., Colozza, G., Moriyama, Y., Benitez, M. D. J., … De Robertis, E. M. (2017). Spemann organizer transcriptome induction by early beta-catenin, Wnt,

Nodal, and Siamois signals in *Xenopus laevis*. *Proceedings of the National Academy of Sciences*, *114*(15), E3081–E3090. http://doi.org/10.1073/pnas.1700766114

Duan, J., Xia, C., Zhao, G., Jia, J., & Kong, X. (2012). Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. *BMC Genomics*, *13*(1), 392. http://doi.org/10.1186/1471-2164-13-392

Glover, N. M., Redestig, H., & Dessimoz, C. (2016). Homoeologs: What Are They and How Do We Infer Them? *Trends in Plant Science*, *21*(7), 609–621. http://doi.org/10.1016/j.tplants.2016.02.005

Grant, V. P. (1971). *Plant Speciation* (1st Editio). New York: Columbia University Press. http://doi.org/0231032080

Guo, X., & Wang, X. (2009). Signaling cross-talk between TGF- β / BMP and other path- ways, 71–88. http://doi.org/10.1038/cr.2008.302

Harland, R. M., & Grainger, R. M. (2011). Xenopus research: metamorphosed by genetics and genomics. *Trends in Genetics*, *27*, 507–515. http://doi.org/10.1016/j.tig.2011.08.003

Hendrickx, M., Van, X. H., & Leyns, L. (2009). Anterior-posterior patterning of neural differentiated embryonic stem cells by canonical Wnts, Fgfs, Bmp4 and their respective antagonists. *Development Growth and Differentiation*, *51*(8), 687–698. http://doi.org/10.1111/j.1440-169X.2009.01128.x

Ho, J. W. K., Stefani, M., Dos Remedios, C. G., & Charleston, M. A. (2008). Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, *24*(13), 390–398. http://doi.org/10.1093/bioinformatics/btn142

Jackson, S., & Chen, Z. J. (2011). Genomic and Expression Plasticity of Polyploidy Scott, *13*(2), 153–159. http://doi.org/10.1016/j.pbi.2009.11.004.Genomic

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics*, *33*(3S), 245–254. http://doi.org/10.1038/ng1089

Joung, J. K., & Sander, J. D. (2013). TALENs: a widely applicable technology for targeted genome editing. *Nature Review of Molecular and Cellular Biology*, *14*(1), 49–55. http://doi.org/10.1038/nrm3486.TALENs

Kjolby, R. A. S., & Harland, R. M. (2017). Genome-wide identification of Wnt/β-catenin transcriptional targets during Xenopus gastrulation. *Developmental Biology*, *426*(2), 165–

175. http://doi.org/10.1016/j.ydbio.2016.03.021

Kondo, M., Yamamoto, T., Takahashi, S., & Taira, M. (2017). Comprehensive analyses of hox gene expression in Xenopus laevis embryos and adult tissues. http://doi.org/10.1111/dgd.12382

Krasileva, K. V, Buffalo, V., Bailey, P., Pearce, S., Ayling, S., Tabbita, F., … Dubcovsky, J. (2013). Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biology*, *14*(6), R66. http://doi.org/10.1186/gb-2013-14-6-r66

Lasky, J. L., & Wu, H. (2005). Notch Signaling, Brain Development, and Human Disease. *Pediatric Research*, *57*(5), 104R–109R. http://doi.org/10.1203/01.PDR.0000159632.70510.3D

Ledford, K. L., Martinez-De Luna, R. I., Theisen, M. A., Rawlins, K. D., Viczian, A. S., & Zuber, M. E. (2017). Distinct cis-acting regions control six6 expression during eye field and optic cup stages of eye formation. *Developmental Biology*, *426*(2), 418–428. http://doi.org/10.1016/j.ydbio.2017.04.003

Lever, E., & Sheer, D. (2010). Dominance and gene dosage balance in health and disease: why levels matter! *The Journal of Pathology*, *220*(September), 114–125. http://doi.org/10.1002/path

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Matthew, P., Leong, J. S., … Vik, J. O. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, *533*(6020), 200–205. http://doi.org/10.1038/nature17164

Liu, Q., & Markatou, M. (2016). Evaluation of Methods in Removing Batch Effects on RNA-seq Data. *Infectious Diseases and Translational Medicine*, *2*(1), 3–9. http://doi.org/10.11979/idtm.201601002

Love, M. I., Wolfgang, H., & Andrers, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *14*(4), R36. http://doi.org/10.1186/gb-2013-14-4-r36

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics*, *92*(1), 155–161. http://doi.org/10.1007/s12041-013-0212-8

Mason, A. S., & Pires, J. C. (2015). Unreduced gametes: Meiotic mishap or evolutionary mechanism? *Trends in Genetics*, *31*(1), 5–10. http://doi.org/10.1016/j.tig.2014.09.011

Michiue, T., Yamamoto, T., Yasuoka, Y., Goto, T., Ikeda, T., Nagura, K., … Kinoshita, T. (2017). High variability of expression profiles of homeologous genes for Wnt, Hh, Notch, and Hippo signaling pathways in Xenopus laevis. *Developmental Biology*. http://doi.org/10.1016/j.ydbio.2016.12.006

Nakade, S., Sakuma, T., Sakane, Y., Hara, Y., Kurabayashi, A., Kashiwagi, K., … Obara, M. (2015). Homeolog-specific targeted mutagenesis in Xenopus laevis using TALENs. *In Vitro Cellular and Developmental Biology - Animal*, *51*(9), 879–884. http://doi.org/10.1007/s11626-015-9912-0

Nakasugi, K., Crowhurst, R., Bally, J., & Waterhouse, P. (2014). Combining transcriptome assemblies from multiple de novoassemblers in the allo-tetraploid plant nicotiana benthamiana. *PLoS ONE*, *9*(3). http://doi.org/10.1371/journal.pone.0091776

Nieuwkoop, P. D., & Faber, J. (1994). Normal Table of Xenopus laevis (Daudin), 418.

Ochi, H., Kawaguchi, A., Tanouchi, M., Suzuki, N., Kumada, T., Iwata, Y., & Ogino, H. (2017). Co-accumulation of cis-regulatory and coding mutations during the pseudogenization of the Xenopus laevis homoeologs six6.L and six6.S. *Developmental Biology*. http://doi.org/10.1016/j.ydbio.2017.05.004

Ochi, H., Suzuki, N., Kawaguchi, A., & Ogino, H. (2017). Asymmetrically reduced expression of hand1 homeologs involving a single nucleotide substitution in a cis -regulatory element. *Developmental Biology*, (March), 1–9. http://doi.org/10.1016/j.ydbio.2017.03.021

Otto, S. P. (2007). The Evolutionary Consequences of Polyploidy. *Cell*, *131*(3), 452–462. http://doi.org/10.1016/j.cell.2007.10.022

Peshkin, L., Wühr, M., Pearl, E., Haas, W., Freeman, R. M., Gerhart, J. C., … Kirschner, M. W. (2015). On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. *Developmental Cell*, *35*(3), 383–394. http://doi.org/10.1016/j.devcel.2015.10.010

Pursglove, S. E., & Mackay, J. P. (2005). CSL: A notch above the rest. *International Journal of Biochemistry and Cell Biology*, *37*(12), 2472–2477. http://doi.org/10.1016/j.biocel.2005.06.013

Ran, D., & Daye, Z. J. (2017). Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. *Nucleic Acids Research*, *45*(13). http://doi.org/10.1093/nar/gkx456

Riddiford, N., & Schlosser, G. (2017). Six1 and Eya1 both promote and arrest neuronal differentiation by activating multiple Notch pathway genes. *Developmental Biology*, *431*(2), 152–167. http://doi.org/10.1016/j.ydbio.2017.09.027

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. http://doi.org/10.1093/nar/gkv007

Session, A. M., Uno, Y., Kwon, T., Chapman, J. A., Toyoda, A., Takahashi, S., … Daniel S. Rokhsar. (2016). Genome evolution in the allotetraploid frog Xenopus laevis. *NATURE*, *538*, 26.

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, *6*(7). http://doi.org/10.1371/journal.pone.0021800

Suzuki, A., Yoshida, H., Heeringen, S. J. Van, Takebayashi-suzuki, K., Jan, G., Veenstra, C., & Taira, M. (2016). Genomic organization and modulation of gene expression of the TGF- β and FGF pathways in the allotetraploid frog Xenopus laevis $. *Developmental Biology*. http://doi.org/10.1016/j.ydbio.2016.09.016

Wagner, A., & Wright, J. (2007). Alternative routes and mutational robustness in complex regulatory networks, *88*, 163–172. http://doi.org/10.1016/j.biosystems.2006.06.002

Watanabe, M., Yasuoka, Y., Mawaribuchi, S., Kuretani, A., Ito, M., Kondo, M., … Taira, M. (2016). Conservatism and variability of gene expression profiles among homeologous transcription factors in Xenopus laevis. *Developmental Biology*, (January), 1–24. http://doi.org/10.1016/j.ydbio.2016.09.017

Wendel, J. F. (2015). The wondrous cycles of polyploidy in plants. *American Journal of Botany*, *102*(11), 1753–1756. http://doi.org/10.3732/ajb.1500320

Winkler, H. (1916). Uber die experimentelle Erzeugung von Pflanzen mit abweichenden Chromosomenzahlen. *Botanik, Bd*, *8*, 417–531.

Wu, H., Wang, C., & Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, *14*(2), 232–243. http://doi.org/10.1093/biostatistics/kxs033

Yang, X., Wang, C., Li, X., Chen, C., Tian, Z., Wang, Y., & Ji, W. (2015). Development and molecular cytogenetic identification of a novel wheat-Leymus mollis Lm#7Ns (7D) disomic substitution line with stripe rust resistance. *PLoS ONE*, *10*(10), 1–14.

http://doi.org/10.1371/journal.pone.0140227

Zhou, X., Lindsay, H., & Robinson, M. D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research*, *42*(11). http://doi.org/10.1093/nar/gku310